TUM School of Computation, Information and Technology
Technische Universität München

Mobile User Behavioral Modelling

Leonardo Tonetto

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung eines

Doktors der Ingenieurwissenschaften (Dr.-Ing.)

genehmigten Dissertation.

Vorsitz:

Prof. Dr.-Ing. Klaus Diepold

Prüfende der Dissertation:

1. Prof. Dr.-Ing. Jörg Ott
2. Prof. Gunnar Karlsson, Ph.D.
3. Assoc. Prof. Dr. Karin Anna Hummel

Die Dissertation wurde am 29.02.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 28.11.2024 angenommen.

Acknowledgement

This was a long and challenging journey, but I was fortunate to have worked with incredibly talented people who made it both possible and enjoyable. I am deeply grateful for the trust and support I received from my supervisor, Prof. Jörg Ott, as well as the invaluable guidance and insightful discussions from Prof. Aaron Yi Ding and Prof. Nitinder Mohan. I also want to express my sincere appreciation to all my co-authors, whose contributions were essential to this work. Finally, I extend my heartfelt thanks to my family, Addie, and my close friends and colleagues for their unconditional support, encouragement during difficult moments, and joyful celebrations along the way.

To me, there is nothing one can dream that cannot be achieved.

Abstract

This thesis explores the intersection of human behavior and mobile device usage, focusing on how devices can sense human behavior and might influence actions. Through analysis of various sources of data, we investigate the relationship between online activities, such as browsing and gaming, and mobility patterns, highlighting their impacts on each other. Our findings underscore the importance of privacy and ethics in behavioral research and raise concerns about potential cybersecurity risks. We propose mathematical models to quantify these influences and discuss implications for network protocols, application design, and epidemic spread prediction. Additionally, we examine crowd mobility using Wi-Fi and Bluetooth data, revealing privacy risks and suggesting alternative methods for studying human mobility while preserving individual privacy. Our work emphasizes the need for ethical considerations in research practices and proposes open questions for future exploration, including less intrusive data capture methods, data control for users, and effective communication of research findings while respecting privacy.

Zusammenfassung

In dieser Arbeit wird die Schnittstelle zwischen menschlichem Verhalten und der Nutzung mobiler Geräte untersucht, wobei der Schwerpunkt auf der Frage liegt, wie Geräte menschliches Verhalten erkennen und Handlungen beeinflussen können. Durch die Analyse verschiedener Datenquellen untersuchen wir die Beziehung zwischen Online-Aktivitäten, wie Surfen und Spielen, und Mobilitätsmustern und zeigen ihre gegenseitigen Auswirkungen auf. Unsere Ergebnisse unterstreichen die Bedeutung des Datenschutzes und der Ethik in der Verhaltensforschung und geben Anlass zur Besorgnis über potenzielle Risiken für die Cybersicherheit. Wir schlagen mathematische Modelle vor, um diese Einflüsse zu quantifizieren, und diskutieren die Auswirkungen auf Netzwerkprotokolle, Anwendungsdesign und die Vorhersage der epidemischen Ausbreitung. Darüber hinaus untersuchen wir die Mobilität von Menschenmengen anhand von Wi-Fi- und Bluetooth-Daten, zeigen Risiken für die Privatsphäre auf und schlagen alternative Methoden zur Untersuchung der menschlichen Mobilität unter Wahrung der individuellen Privatsphäre vor. Unsere Arbeit unterstreicht die Notwendigkeit ethischer Überlegungen in der Forschungspraxis und schlägt offene Fragen für die künftige Erforschung vor, darunter weniger aufdringliche Datenerfassungsmethoden, Datenkontrolle für die Nutzer und effektive Kommunikation von Forschungsergebnissen unter Wahrung der Privatsphäre.

Contents

List of Figures

List of Tables

Acronyms

ACS	Absorption Cross Section.
AP	Access Point.
API	Application Programming Interface.
AWDL	Apple Wireless Direct Link.
BLE	Bluetooth Low Energy.
BSSID	Basic Service Set Identifier.
BT	Bluetooth.
CCPA	California Consumer Privacy Act.
CCR	Credit Card Records.
CDF	Cumulative Density Function.
CDR	Call Detail Records.
CFS	Correlation Feature Selection.
CNN	Convolutional Neural Network.
CPR	Control Panel Records.
CS	Computer Science.
CSI	Channel State Information.
CTS	Clear to Send.
ERB	Ethical Review Board.
EU	European Union.
FSPL	Free-space path loss.
GDPR	General Data Protection Regulation.
GPS	Global Positioning System.
IP	Internet Protocol.
IRB	Institutional Review Board.
LDPL	Log-Distance Path Loss.
LSTM	Long Short-Term Memory.
MAC	Media Access Control.

NDA	Non-Disclosure Agreement.
OS	Operating System.
OSN	Online Social Networks.
OUI	Organizationally Unique Identifier.
PDF	Probability Density Function.
PII	Personal Identifiable Information.
PL	Path Loss.
PoC	Proof of Concept.
POI	Point Of Interest.
RFID	Radio-frequency identification.
RSS	Received Signal Strength.
RTS	Request to Send.
SSID	Service Set Identifier.
SVM	Support Vector Machine.
TUD	Time Until Done.
UID	User Identifier.
XDR	eXtended Detail Records.

List of Publications

This thesis consists of an overview and of the following publications. All publications are subject to a full peer-review process.

[2] Alipour, B., **Tonetto, L.**, Ding, A. Y., Ketabi, R., Ott, J., and Helmy, A. (2018). Flutes vs. Cellos: Analyzing mobility-traffic correlations in large wlan traces. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications* (pp. 1637-1645).

[3] Alipour, B., **Tonetto, L.**, Ketabi, R., Yi Ding, A., Ott, J., and Helmy, A. (2019). Where are you going next? A practical multi-dimensional look at mobility prediction. In *Proceedings of the 22nd International ACM Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems* (pp. 5-12).

[4] **Tonetto, L.**, Untersperger, M., and Ott, J. (2019). Towards exploiting Wi-Fi signals from low density infrastructure for crowd estimation. In *Proceedings of the 14th Workshop on Challenged Networks* (pp. 27-32).

[5] **Tonetto, L.**, Lagerspetz, E., Ding, A. Y., Ott, J., Tarkoma, S., and Nurmi, P. (2021). The mobility laws of location-based games. In *EPJ Data Science*, 10(1), 10.

[6] **Tonetto, L.**, Adikari, M., Mohan, N., Ding, A. Y., and Ott, J. (2022). Contact duration: Intricacies of human mobility. In *Online Social Networks and Media* (28, 100196).

[1] **Tonetto, L.**, Carrara, A., Ding, A. Y., and Ott, J. (2022). Where is my tag? unveiling alternative uses of the Apple FindMy service. In *2022 IEEE 23rd International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)* (pp. 396-405).

[7] Kister, P., and **Tonetto, L.** (2023). On the importance of structural equivalence in temporal networks for epidemic forecasting. In *Nature Scientific Reports* (13(1), 866).

1 Introduction

Understanding human behavior, especially mobility, has been a many-decades-long quest for researchers. To conduct such studies, data about people's movements had to be collected, initially through analog records (*e.g.*, logbooks and agendas) and, more recently, through a vast range of electronic sources. The variety and growth of these digitally sampled records reached its current apex with the introduction and popularization of smartphones, in the early 2000s. These always-connected devices allowed not only users to communicate but also empowered billions of people with numerous sensors capable of accurately describing their behavior. While this rapid increase in data availability facilitated research it also exposed users' individual privacy to unprecedented levels of risk. In turn, researchers had the ability to study aspects of human mobility which were not possible to address in the past. Furthermore, this vast availability of mobility data, where some are private, created two important challenges, among many others: (**C1**) how to analyze such big sets; and (**C2**) how to do so while protecting the subjects' privacy.

To address C1, various big data approaches were introduced, which mostly consist of breaking down the desired problem into smaller, more tractable, sub-problems. These methods allow us to model and better understand the problems being studied, as well as create accurate predictions of unseen cases while often being comparable and reproducible. To address C2, among the extensive literature available, methods either require some form of anonymized data (*e.g.*, user identifiers are replaced, anonymity sets, *pseudo*-synthetic data) or limit their results to never (or only temporarily) require any form of identification. While both of these challenges have been extensively addressed in the past, the ever evolving datasets (*i.e.*, volume and variety of sources) on human mobility as well as the constantly changing understanding of the general population, regulations and ethical guidelines towards individual privacy foster the need for more research.

The aforementioned rapid expansion and popularization of mobile smart devices, such as phones, tablets and laptops was made possible, at least in part, by modern wireless communication technologies. These technologies include those used for telecommunications, such as mobile phone networks (3/4/5G) and Wi-Fi, as well as very short range networks such as Bluetooth and RFID. These various alternatives for connectivity enabled mobile broadband access to the Internet along with seamless integration between nearby devices and their peripherals. Finally, it was through the logs and fingerprints of this expanded connectedness that large amounts of human behavioral data could be obtained. While such data may have originally been sampled to improve the very underlying system they were taken from, researchers realized its potential to study how people move, how they use the internet, and how people interact physically and virtu-

ally [8]. It is worth noting that other sources of mobility data have been used, such as public transport tickets, but this work has focused on mobile devices as its main source of information about subjects.

Given the discussion above, we list three main problems that are worthwhile expanding, and which we will present as central pieces of this thesis:

- Understanding human mobility. To model and explain what affects people's macroscopic movements is of paramount importance to better systems' design, urban planning and epidemic mitigation.
- How to study mobility while preserving the individual privacy of subjects. By either not requiring any User Identifier (UID) or only keeping a UID for a desired computation task (*e.g.*, count estimate at a given time).
- Privacy and ethics. When systems evolve too fast it may open subjects to unforeseeable risks. This is valid both for systems providing new/better access to different resources and research being done under little scrutiny.

We look forward to a future where data on human behavior are available to studies, from users who do not have second thoughts about sharing their data because of privacy as that is safely handled. In this thesis, we investigate different aspects of human mobility, including its understanding through empirical and predictive models as well as the privacy and ethical risks involved with such data. We present results which help us understand how mobility and network utilization are intertwined and influenced by the applications being used as well as devices being used. Furthermore, we present results of models between human mobility and expected contact duration along with evidence that stay duration at different places may be associated with the type of activity performed at each location. Utilizing passive measurements, we present results on how different aspects of human mobility may be studied without the need for any UID. Different from the previous set of results, these privacy preserving alternatives focus on crowd assessments. Moreover, we present results on how mobile users' privacy might have been at risk from Bluetooth finders. We also present a survey of how privacy and ethics have been addressed in existing studies on human mobility.

1.1 Problem statement

In this thesis, we address two different but interconnected problems regarding human mobility: (1) the understanding of how people move and how these movements influence and are influenced by other things, and (2) the privacy and ethical implications of the data being collected from a growing number of sources.

Given the problems listed above, these 3 research questions will set the main objectives of this thesis and guide us through its contributions:

- **RQ1: How to establish and study the relationship between human mobility and mobile device usage?** While mobile devices become more popular

and always connected, their usage is intrinsically related to mobility as people get access to various online services, such as maps, media and games. Understanding this relationship is of great importance for the improvement of mobile connectivity as well as the monitoring and decision-making related to mobility.

- **RQ2: How can we effectively study certain aspects of human mobility without compromising subjects' privacy?** Individual mobility data capture a vast amount of personal information about subjects, often beyond what the data were originally sampled for. This may raise privacy issues and, more recently, break the law in various countries. Therefore, privacy-preserving methods that do not require a permanent storage of identifiers is required so that infrastructure providers can better plan and react to how people move.
- **RQ3: Are mobile tracking devices privacy preserving as advertised?** Mobile tracking devices, originally designed to help find lost objects while preserving the privacy of its users, may disclose more information than required. In that case, private information may be disclosed as well as this system may be used to unwittingly transport data between two endpoints.

We begin by establishing the relationship between human mobility and mobile device usage. The popularization of such devices, along with their continuous connectivity to the Internet and other nearby devices have shaped how we move and how we interact. Therefore, it is of relevant to better understand how we use different devices, given their size and portability as well as how mobility and network traffic are related, which we address in this thesis. Additionally, we address the problem of how predictable a mobile user is. In another work, we study how contact duration and stay duration relate to the mobility of mobile users. Similarly, we also look at how an online game with a strong mobility component influences players' movements when playing it. It is worth noting that these results were obtained through analysis of large *anonymized* individual mobility data.

The use of such individual data, even when anonymized, may raise concerns regarding privacy. Therefore, approaches to study aspects of human mobility that do not require identifiers is of great relevance. Recent developments in state regulations as well as comprehensive ethical guidelines regarding the protection of individual privacy further challenge the study of human behavior. To address these challenges, we present studies that use passive measurements and are able to infer various aspects of a crowd, without requiring the persistent storage of any UID. It is worth noting that these measurements are based on wireless signals, part of systems that were not originally designed for such crowd assessments.

The rapid expansion and pervasiveness of mobile connectivity created not only research opportunities to study human mobility but also introduced new security, privacy and ethics challenges. Extended connectivity between devices increases not only the functionalities of a system but also widens the attack surface a mobile device user could be exposed to. To that end, we present results from a study where we identify a series of potential security risks mobile users could be exposed when using Bluetooth object

trackers. These threats we identified could expose private information about nearby phone users. In addition to potential security risks to an individual privacy through their mobile devices, data previously collected for research studies may also raise privacy and ethical concerns. To better understand this issue, we present a literature survey, reviewing over 20 years of articles that studied human mobility using individual mobility traces. This review focuses on how personal data were used and how the work was communicated in the published article.

1.2 Research Methodology Overview

To study mobility, we processed and analyzed large individual traces which were sampled from a wide variety of sources, such as Wi-Fi network traces, Online Social Networks (OSN), phone app utilization and controlled experiments using Global Positioning System (GPS). To capture unique properties of these traces, we modeled different behaviors fitting empirical distribution functions, such as powerlaw and exponential. We investigated the functioning mechanisms of a mobile tracking system in order to show how private information about users could leak and how this system could be used to exfiltrate data, inadvertently. All analyses of private data done for this thesis strictly followed relevant regulations for data protection. No single individual was studied in isolation, and no attempts were made to de-anonymize the available data. Chronologically, the studies done in the later part of this thesis we focused on crowd mobility, where all data being collected were aggregated at the time of collection. This shift aimed at addressing evolving data privacy considerations, including the implementation of regulations like GDPR.

1.3 Approach

The studies we present in this thesis were the result of analysis of mobility data and a review of the existing literature on related topics. In some cases, the data we used were already available, and were sampled prior to the study. In the other cases, the study we did included data captured for the exact purpose of that research. These different approaches resulted in a series of heterogeneous datasets, each capturing different aspects of mobility at various granularity levels, requiring a set of robust analyzes for reliable results. We logically separate these results into two groups: (1) **Understanding human mobility through mobile devices**, in which we address the more fundamental questions about human mobility as seen through the usage of mobile phones. (2) **Understanding human mobility while preserving privacy**, that in contrast to (1), we use approaches in which user identifiers are not required, except for the validation of the results, as well as a case where we study the leakage of private information on mobile systems. Next, is a short summary of how this was done and how it will be presented in this thesis.

Understanding human mobility through mobile devices – Chapter 3 We study differences between easily portable devices, such as smartphones, and bulkier ones, such as laptops, in terms of mobility and network traffic. For that, we propose the FLAMeS framework to analyze large mobility traces along with network traffic data. Through this framework, we unveil fundamental characteristics of human mobility and how they relate to network utilization by each of the studied devices. This investigation was done using a large Wi-Fi trace, consisting of 100s of thousands of devices, which we classified according to their portability (*i.e.*, mobile phones and laptops). Using the same dataset from the previous point, we studied how predictable human mobility could be in a university campus scenario through entropy. The work also includes a series of predictive models which validate the observations made using the information theory approach. Furthermore, we analyze data from a power monitoring app (*i.e.*, Carat [9]) along with geotagged posts from an OSN (*i.e.*, Twitter), we study how players of an online game (*i.e.*, PokémonGo) changed their movements during the months they played the game. The selected game has a strong mobility component, where users were required to move large distances in order to achieve goals in the game. For this study, empirical distribution fitting was used in order to study how movements changed over time, especially when comparing periods before and after the introduction of the game, with a control group to validate the results. Additionally, using a similar approach, but with fine granular mobility data from recruited subjects, we study how the expected duration of contacts between nearby mobile users and the expected duration a person will stay at a place related to their mobility. For this, we augment the original dataset with metadata about the places being visited by each subject which allowed us to correlate visit duration and location labels. Finally, our observations on mobility and contacts were further supported by results from a separate study where we examine the impact of network structure on disease prediction. Taken all together, these studies reveal intricacies between human mobility and other behaviors, such as phone usage, and either expand our understanding or reinforce previously made observations on how people move.

Understanding human mobility while preserving privacy – Chapter 4 The goal of this part is to investigate methods which enable the study of human mobility while disclosing as little personal information as possible, especially without the necessity to permanently store any UID. For that, we present results of a study in which we assess the occupancy of a confined area through changes in the signal strength of received nearby signals. This allows us to infer relative counts (or density) in a monitored area. Furthermore, following a similar approach to repurposing a system to aid in crowd monitoring, we present results of using reports from a Bluetooth tracking system to estimate different aspects of a crowd. Such tracking systems rely on a network of sensing devices (*i.e.*, phones) to help point the location of *lost* devices. The reports sent are anonymized and end-to-end encrypted, establishing a secure approach to estimate crowd density and flow. However, the study of such system revealed potential privacy risks to their users, which are included in the next set of results. Taken all together, the observations made with these studies point to potential alternatives to study human movements while attempting

to protect subjects' privacy. We note that, these privacy issues raise not only the need for more secure systems but ethical considerations to be taken while analyzing such sensitive data. While human mobility studies may include various sources of (bigger) data to ensure their validity, not enough has been done to ensure the privacy of subjects. This ethical aspect while not addressed in this thesis, begs for further studies.

Additionally, **Chapter 2** revises fundamental methods relevant for different chapters, and finally **Chapter 5** concludes this thesis with a summary of findings and discussions about common topics covered.

2 Background

In this chapter, we present the fundamental definitions of the methods used in this thesis. We begin by discussing the various possible sources of data on human mobility (*i.e.*, traces). Next, we discuss a series of models we later use to study and explain human mobility, including empirical distributions and entropy.

2.1 Data Sources

To better understand human mobility, previously collected *traces* are a relevant source of information. They often capture data about movement of the subjects, including a location and a timestamp. These traces may allow us to (1) reconstruct the trajectories of an individual (or group of persons), (2) model features of their visits, such as regularity and duration, and (3) model features of the paths taken, such as distances traveled. We now review the most commonly used sources of mobility data based on mobile device usage.

Note, however, that this section does not include other commonly used data sources for research studies. We do not include *simulations*, which as a source for data are only as good as the model they are based on, *i.e.*, may not be good representations of certain aspects of reality.

2.1.1 Communication Infrastructure-based Data

Communication infrastructure-based data became ubiquitous in mobility research with the popularization of connected devices in the early 2000s. Such sources rely on existing communication infrastructure to sense the presence of a subject, *e.g.*, when a mobile phone connects to a cell tower. Their presence is inferred by records (or logs) created when a subject's mobile device communicates with a point of access, which could be of a Wi-Fi network or of a mobile cellular network [8]. While Wi-Fi setups are often limited to confined areas, such as companies or university campuses [10, 2], they offer building- and room-level accuracy for their locations. Mobile cellular networks, on the other hand, may cover entire countries but have their accuracy between 100s of meters to 10s of kilometers [11] with Call Detail Records (CDR) in most modern setups (*i.e.*, 3/4/5[1]). In both cases, the availability of location records may be a function of the activity of the user, such as when an phone call or SMS message is received or sent, or a device associates to an access point upon arriving to a new area. Additionally,

[1]Note that, although 5G promises much denser deployments, its current density is similar to that of 4G in most networks [12]

certain network settings allow location records to be captured whenever *any* data or signaling events happen between the network and a subject's device (*e.g.*, eXtended Detail Records (XDR) and Control Panel Records (CPR) [13]). For such, subjects often agree on *terms and services* for the use of the offered infrastructure, and the collection of location data can be seen as a byproduct of this interaction. Alternatively, Credit Card Records (CCR) can also provide rich information about its user's whereabouts whenever a card is used for payment at a physical store [14].

While the data discussed above may contain great amounts of information about each subject, they are often owned and controlled by infrastructure providers, such as telecom operators. Additionally, current data regulations in many countries impose a series of limitations to the sharing of such data, restricting mobility research from using such data. Alternatively, passive measurements are capable of capturing data from signals used for communication, such as Wi-Fi and Bluetooth. Nearby devices can be detected through signals that were originally standardized to aid with discoverability in both technologies. The Wi-Fi standard defines different frame types, among which *management* frames are responsible for access point advertising its existence, for mobile devices to actively search for access points (*e.g.*, hidden networks), and to support connectivity between devices. One relevant type of Wi-Fi management frame are *probe requests*, which help devices actively probe for nearby access points. These frames may contain, among others, an identifier for the issuing mobile device. When passively intercepted, these probe requests can register the physical presence of a nearby device, at a specific time and location. Finally, when measuring such signals over a large time frame and possibly multiple locations, the mobility of a device can be reconstructed. That is, the sequence of events (*i.e.*, visit to a location) and timestamps of a UID. A similar approach can be used for Bluetooth (BT) signals, where BT devices use a similar protocol to advertise their availability or search for other devices [15]. Note, however, that modern versions of mobile operating systems (*e.g.*, Android and iOS) increasingly use randomized UID in order to protect the privacy of their users [16].

2.1.2 Designed Experiments

Another class of human mobility data source used in research originates from deliberately designed experiments, in which a phone app is installed, or pre-configured devices are distributed among subjects (*e.g.*, [17, 18]). These types of efforts typically provide the highest level of flexibility and uniformity in how data are sampled, at times also providing data from other non-location sensors, such as accelerometers. This higher stability is the result of pre-defined configurations set by the experiment, resulting in traces that are sampled at regular intervals (*i.e.*, after a certain amount of time or distance traveled). These extra readings enable a higher accuracy in segmenting events, such as the duration of stops. However, as these studies are based on recruiting people, and given its costly setup and lack of secondary benefits for subjects (such as Internet access through a Wi-Fi access point), cohorts are limited in their size when compared to communication infrastructure based efforts [18]. Location data collected through these studies often include continuous GPS coordinates, BT and Wi-Fi scans of nearby devices.

Designed experiments may also include passive measurements, often with the use of BT or Radio-frequency identification (RFID) tags [15]. In such studies, subjects are given such tags with an associated UID, which are probed at regular intervals, establishing a reliable source for location and proximity between subjects over time.

2.1.3 Internet-Based Services

Another essential class of service providers that capture location data are those enabling the exchange of information on the Internet. Web services, such as OSN [19], search engines [20] and map services [20] provide online users with services while also logging their physical location, for example by means of geotagged posts or geolocating devices based on their IP address [21, 22]. Unlike the sources previously discussed, the sampling of location records by web services is highly dependent on how often a subject interacts with those services, leading to a skewed availability of data per device. Alternatively, location data may be captured in the background but only if allowed by the user (*e.g.*, [23, 24]). Additionally, these web services can also capture extra features, such as the content of what is being searched or posted as well as the social graph of their users, which can be used to further enrich any analysis being made. Similar to the communication infrastructure, subjects agree to *terms and conditions* for that service that states their location data will be logged and may (or may not) be used for further studies.

Before recent updates on mobile operating systems, passive data collection of location data was possible by any installed application. This allowed researchers to crowdsource human mobility information, without the consent of the participants, and without any UID.

2.1.4 Public Transport Infrastructure

SmartCards have replaced old paper-based ticketing system in most modern public transportation systems in large metropolitan areas [25]. They provide an integrated and automated way for passengers to pay for transport rides as well as manage different pricing schemes (*e.g.*, senior citizens or students discounts). Users are required to present their smart cards before starting a ride, for example when entering a subway station or boarding a bus, and, in some cases, the same is expected when alighting. In this way, the system records the timestamps of discrete location points a subject has been. As humans tend to produce repeatable mobility patterns [26], location data from public transport tends to be consistent and homogeneous through time, at least until a global pandemic changes how people move. Similar to all infrastructure-based sources, subjects agree to the *terms and services* of using a smart card for their corresponding transport system. Additionally, smart cards have also been used to trace the behavior of students in a university campus, logging various activities and services used by students [27]. Other examples of mobility data gathered from public transport are shared bikes [28], and taxis [29].

2.1.5 Others

There are other sources of mobility data that have not been used in this thesis, either for the privacy concerns raised with their use or for their unreliability.

Image/Camera-based Current methods in computer vision allow for highly accurate identification and tracking of subjects [15]. These methods, however, raise privacy concerns as not only identifiers might be kept, but also the depiction of facial and body features, easily identifiable by a human subject.

Survey/Questionnaire-based This alternative source of mobility data is still currently used, especially to study events or phenomena that could not be predicted, such as the use of a new phone app or curfew measures during a pandemic. The information contained in these logs is often biased, based on what subjects are able to remember to what is being asked. Records are stored a posteriori and might include a detailed agenda of events and timestamps or a questionnaire capturing the desired changes [30].

Given these different types of datasets, their availability and information captured about mobility, we now turn to methods with which insights can be extracted from these data.

2.2 Modelling Mobility

To model a process or action is to describe a phenomenon in such a way that it can be reproduced, and likely explained. A model might be a mathematical formulation or an algorithm, capturing one or more aspects of the phenomenon being studied. Before the aforementioned increase in availability of mobility data, synthetic mobility models were used to represent the movement of individuals and study their consequences [31, 32]. Such synthetic models include *random walks* and *random waypoints* [32], in which mobile nodes move freely in space, choosing their next steps randomly and independently of each other. While simulations using such models can scale to large numbers of nodes and various properties can be mathematically defined (*e.g.*, [33]), these models do not capture essential characteristics of human mobility such as short and long-term regularity [19, 26, 34], long-tailed distributions of stay duration or displacements [35, 36], or bounds in the number of visited locations over time [37]. To address some of these shortcomings later random models were proposed, such as *Lévy-Flights* [11] in which hops follow a long-tailed (powerlaw) distribution or the *working day movement model* which splits a working day into different periods and reproduces regular movements between relevant places for a node (*e.g.*, home and work) [38]. These models enabled large scale evaluations of mobile network protocols while addressing a selected number of limitations found on early random models. However, trace-based analysis is required to better understand how humans move across space as real data capture a combination of nuances not present in any random model thus far and ultimately, the above models are built and validated by using traces.

Trace-based analysis of human mobility [8] has been proven to be relevant for not only network protocol design but also urban planning [39], health risk prevention [40], social sciences studies [41], and, more recently, epidemic management [40]. The analysis of these traces can reveal new insights about human mobility, and validate, reinforce or refuse existing ones. This is often done through graphical representation of the data collected (*e.g.*, [42]), statistical aggregates of observed metrics (*e.g.*, [8]) or through empirical models (*e.g.*, [11]). In this thesis, we utilize all three approaches to present and discuss observations and findings from the studies we conducted.

Given the variety and complexity of the empirical distribution modelling used in this thesis, we now review the functions that we use and the theory behind the method for finding the best fit.

2.2.1 Empirical distribution modelling

While highly parametrized models, such as neural-networks, are applicable to any type of data including hyper-dimensional sets, extensively studied probability distributions are often more *interpretable* (*i.e.*, changes in the distribution of the input data can often be explained by variations in the respective parameters of the modelling functions), *comparable* (*i.e.*, a different set of parameter values or different distributions have fundamental properties that can be compared), and *portable* while preserving the **privacy** of the subjects involved in the study (*i.e.*, models or datasets can be compared with little to no personal identifiable information being shared alongside a scientific publication).

In various analyses in this thesis, we observe and assess the likelihood of the following three long-tailed distributions, which are now presented along with typical implications of observing each one of them. Unlike their exponential counterparts, heavy-tailed distributions are not bounded at any given value, but rather maintain a certain (non-negligible) probability to all values. Furthermore, these probabilities are often proportional to the value assumed by the variable they model, such as ranks in a Zipfian distribution [43]. Note that, while several other probability distribution functions have been proposed in the past, these three are the ones most commonly used to describe human behavior, making it a simpler comparison with related work.

Log-normal The Probability Density Function (PDF) of this function, for a given random variable X for all $x > 0$, is defined by Equation 2.1, with parameters μ (mean or *location*) and σ (standard deviation or *shape*). Intuitively, this distribution describes a Normal distribution for the logarithm of a random variable. This distribution has been used to describe trip length from GPS data [44, 45, 46] and for stop duration [45], for describing the length of textual Internet content [47], and time users spend on individual Internet content without a time component [48] (*e.g.*, images, text).

$$p(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right) \tag{2.1}$$

Weibull The PDF of this distribution function, for a given random variable X for all $x > 0$, is defined by Equation 2.2, with parameters λ (*scale*) and β (*shape*). While λ describes how spread-out the distribution is, β defines whether the tail of the distribution will be exponential (when $\beta > 1$) or long-tailed (when $\beta < 1$). This distribution has been used to describe trip length from Twitter data [19] and from taxi data [49], as well as users behavior on online social networks [50].

$$p(x) = \frac{\beta}{\lambda}\left(\frac{x}{\lambda}\right)^{\beta-1} e^{-\left(\frac{x}{\lambda}\right)^{\beta}} \tag{2.2}$$

Power-Law The PDF of this distribution function, for a given random variable X, is defined by Equation 2.3, with parameters α (*scale*) and $x_{\min}$ where $\alpha > 0$ and $x_{\min} > 0$. This distribution has been extensively used to model various naturally occurring phenomena [51] and is often explained by *preferential-attachment* in a time-evolving network [52]. Power-law models have been extensively used to describe trip length [36, 11], friendship on online social networks [53], and the organization of the Web [54].

$$p(x) = \frac{\alpha - 1}{x_{\min}}\left(\frac{x}{x_{\min}}\right)^{-\alpha} \tag{2.3}$$

2.2.2 Parameters estimation and distribution comparisons

In all the problems we study, we use the maximum-likelihood method proposed by Clauset *et al.* to fit the parameters of these distributions, which provably gives accurate parameter estimates in the limit of large sample sizes [55]. That is, to find the best set of parameters for each of the three PDFs above, we compare the likelihood that each distribution best describes the data we are trying to model. First, the best parameters are found for a distribution by minimizing the distance between the data provided and the PDF used to model it. Effectively this is done using the Kolmogorov-Smirnov test (KS-test) [56], which yields a KS-distance (or KS-statistic) metric between data and model, which can be then minimized. Finally, following the method by Clauset *et al.* [55], based on the KS-statistic and likelihood ratio, we produce a *p-value* which allows us to infer the significance of this comparison (*i.e.*, that it was *not* due to chance). In this thesis, we adopt the common convention that a *p-value* < 0.05 is significant. That is, when comparing how well two distributions describe a set of data, a *p-value* < 0.05 indicates that there is a probability lower than 5% that the best distribution was chosen due to randomness. Therefore, whenever reporting a distribution fit, we compare the goodness-of-fit between the empirical distribution functions discussed above: Log-normal, Weibull, and Power-Law and provide the *p-value* to the comparison between the two best options.

Note that the goodness-of-fit described above finds the best set of parameters for each candidate distribution and compares which of the candidates best fits the data. That is, this approach can use any arbitrary number of candidate distributions, yielding the likelihood any of them best fits the data alongside a *p-value* for validation. Alternatively,

the graphing (*i.e.*, plotting) of each candidate and visually comparing the distributions have been used [55] which may produce inconsistent or even subjective results.

2.2.3 Mobility metrics

To study human mobility, various metrics may be used with varying degrees of complexity. Simple numerical metrics such as the number of visited locations, and rank-based metrics such as top-visited locations captured general descriptive observations. These however, do not include a temporal or spatial dimension to the analysis of mobility, which may be achieved with metrics such as flight length, which describes the distance traveled between two points in space or time, or diameter, which captures the largest distance a device may travel during an observation period. Furthermore, to include a temporal aspect to such metrics, Gonzalez *et al.* [11] devised the radius of gyration for human mobility.

Radius of Gyration (r_g) This metric captures spatial dispersion of a subject's trajectory. This metric is defined by Equation 2.4:

$$r_g(t) = \sqrt{\frac{1}{n_c(t)} \sum_{i=1}^{n_c} (\overrightarrow{r_i} - \overrightarrow{r_{cm}})^2} \tag{2.4}$$

where $\overrightarrow{r_{cm}} = \frac{1}{n_c} \sum_i^{n_c} \overrightarrow{r_i}$ is the centroid of the locations a device visited over a period of time, and $\overrightarrow{r_i}$ is location $i = 1, \ldots, n_c(t)$ until time t.

Isotropy ratio This metric describes the spatial dispersion of a set of trajectories given a common reference frame (e_x, e_y), as proposed by Gonzalez *et al.* [11]. That is, a user with an anisotropic pattern of movements would have its visited points found along one axis e instead of scatter between both axes. While radius of gyration captures the size of the area trajectories are found, isotropy ratio allows us to study the dispersion of these points within this area.

2.2.4 Entropy

The *entropy* of a random variable – also known as *Shannon entropy* [57] – is a measure of the average uncertainty in realizations of this random variable. Consider a single observation of a person's location as generated by a discrete random variable X taking values from a finite set $\mathcal{Z}$ (alphabet) with probability distribution $p(x), x \in \mathcal{Z}$. The entropy of X is given by

$$H(X) = -\sum_{x \in \mathcal{Z}} p(x) \log p(x). \tag{2.5}$$

Throughout this thesis, we will use logarithms to *base 2*.

Analogously, entropy can be associated with a stochastic process $\mathbf{X}$ defined as a family of random variables $\{X_t, t \in T\}$ taking values from the same alphabet and with the same probabilities at each time step. Given the amount of information known about a random

process, we can quantify the uncertainty in predicting future realizations of a time series, generated by this process, with three related entropy metrics.

Firstly, if only the set of possible outcomes is known, that is, the finite alphabet from which the realizations are drawn—in our case the alphabet represents the set of locations that the user visits—predicting the future steps corresponds to choosing one among all, equally-probable locations. It is well known that such uniform probability distribution yields maximum uncertainty—in our notation, maximum entropy $S^{\max}$.

Secondly, if the frequency of visits is observable, but not the order in which the user visited these locations, one can make a more informed guess based on the probabilities of the possible events. Since the temporal dependencies between visits are unobservable in this case, we refer to the corresponding entropy measure as *(temporally) uncorrelated entropy* and denote it by S^{unc}.

Lastly, if the ordered sequence of visits is known, for predicting the next step of a time series one can take all the aforementioned information into account which yields an uncertainty often referred to as the *entropy rate* S^*. For a stationary time-series it can be shown that $S^{\max} \geq S^{\text{unc}} \geq S^* \geq 0$ holds [26], meaning that the uncertainty about the next realization of a random process decreases when more information is gathered about it.

Next, we mathematically describe these three entropy measures.

Maximum Entropy

Consider a random process $\mathbf{X}$ and assume that this process generates, at each of the observed discrete time steps, one of the N_{loc} equally likely events $x \in \mathcal{Z}$. From Equation (2.5), it can be easily seen that the entropy of this process is $\mathbf{H}(X) = \log N_{loc} = S^{\max}$ and that the entropy increases with the number of possible outcomes, that is locations. In terms of prediction, the more locations a user has visited, the higher is the uncertainty about his whereabouts and the next location that he will visit.

Uncorrelated Entropy

Given the probability distribution of the outcomes of the process $\mathbf{X}$, the uncertainty of the *instantaneous* location where the user is can be regarded as the uncorrelated entropy S^{unc}, which is precisely the Shannon entropy from Equation (2.5), $S^{\text{unc}} = -\sum_x p(x) \log p(x)$. Intuitively, if two users visit the same number of locations but one with a uniform distribution and the other user following some skewed distribution, for instance a *powerlaw* distribution, then the uncorrelated entropy of the first user, S^{unc}, will be much greater than the same metric for the second user.

Entropy Rate

The uncertainty in forecasting future observations of a time series can be further reduced by considering temporal correlations among observed samples. For a temporally

ordered random process $\mathbf{X} = \{X_t : t = 1, \cdots, n\}$, we define a sub-series X_i^j as the segment $X_i^j = \{X_i, \cdots, X_j\}, 1 \leq i, j, \leq n$ and denote by $\mathbf{x}_i^j = (x_i, \cdots, x_j)$ one possible realization. The entropy rate of $\mathbf{X}$ is defined as the asymptotic rate at which the entropy of X_1^n changes with n

$$H(\mathbf{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, \cdots, X_n) \tag{2.6}$$

where $H(X_1, \cdots, X_n)$ is the joint entropy of n random variables X_i. For a *stationary* process the limit exists [58, 59] and the entropy rate converges to

$$H(\mathbf{X}) = \lim_{n\to\infty} H\left(X_n | X_1^{n-1}\right) \tag{2.7}$$

where $H\left(X_n | X_1^{n-1}\right)$ is the *conditional entropy*, that is, the entropy of the random variable X_n conditioned on the previous events X_1^{n-1}. Intuitively, if two users visit the same number of locations, where the first user cycles between these places following always the same order whereas the other user visits the places randomly, then after an arbitrarily long observation and using only the knowledge of past events, the uncertainty in guessing the next step of the first user will be smaller than that of the second user.

To summarize, among the three metrics, the *entropy rate* (S^*) holds the promise of giving the most accurate prediction of the next location, since it yields the lowest uncertainty by exploiting the information about the order and the frequency of visits to different locations. Utilizing the entropy rate for prediction, however, has to be implemented with some caution. First, Equations (2.6) and (2.7) hold under the assumption that the underlying stochastic process is stationary. Second, even for a stationary process, one can only obtain an *approximation* of the entropy rate; this approximation depends on the number of observations, as well as on the estimation method. We now further elaborate on different approaches to entropy rate estimation.

Entropy Rate Estimators

The most common methods for estimating the entropy rate are based on data compression algorithms [60] that achieve optimal compression rates by asymptomatically approaching the entropy rate of the underlying process. Among the data compression algorithms, the best known is the Lempel-Ziv (LZ) compression algorithm [61]. Several other approaches have also been proposed: a method based on Context-Tree Weighting (CTW) [62, 58], which is often used for binary sequences, and methods using Burrows-Wheeler transform (BWT) [63, 64].

LZ estimator This entropy rate estimator is based on the calculation of the length of repeating patterns in the data. For a random process $\mathbf{X}$ and its realizations $\mathbf{x}$, let L_i^n be the length of the longest segment $x_i^{i+\ell-1}$ starting at position i and length ℓ which has also appeared in the window x_{i-n}^{i-1} with length n preceding position i [58]. The LZ estimator gives an approximate entropy rate of the process

$$\hat{H}_n(\mathbf{X}) = \mathbf{S}^{\mathrm{lz}} = \left[\frac{1}{n}\sum_{i=2}^{n} \frac{L_i^i}{\log i}\right]^{-1} \tag{2.8}$$

in which the window length n increases as the matching position moves forward, taking the entire memory of the system into account.

BWT estimator The BWT based entropy rate estimator first rearranges the input sequence by means of the Burrows-Wheeler Transform, which preserves the original symbols in the sequence, but generates runs of repeated symbols. This rearrangement can be efficiently computed by constructing a suffix-tree [65] from the input sequence, then collecting all leaf nodes by lexicographically traversing the resulting tree [66]. The estimator then averages the entropy of sub-segments of the transformed input [64]. For this second step, the transformed input is segmented into k chunks $c \in C$ of uniform length $\ell = \sqrt{n}$, where n is the length of the input sequence. The first order distribution within each chunk is estimated from Equation (2.9), where $N_c(x)$ is the number of occurrences of the symbol x in the chunk c. Next, the entropy of each chunk is computed from Equation (2.10), and finally, the BWT entropy rate is estimated by averaging the individual entropies per chunks from Equation (2.11).

$$\hat{q}(x, c) = \frac{N_c(x)}{\sum_{y \in \mathcal{Z}} N_c(y)}, \tag{2.9}$$

$$\log \hat{q}(c) = \sum_{x \in \mathcal{Z}} N_c(x) \log \hat{q}(x, c), \tag{2.10}$$

$$\hat{H}_c(\mathbf{X}) = \mathbf{S}^{\mathrm{bwt}} = -\frac{1}{n} \sum_{c \in C} \log \hat{q}(c). \tag{2.11}$$

This set of steps allows us to compute the Burrows-Wheeler Transform of a given sequence, which has shown to produce more robust estimates than the LZ method [66], *i.e.*, the BWT is better at capturing long- and short-term context than LZ which allows it to be less sensitive to local changes of information.

2.3 Summary

In this chapter, we reviewed commonly used data sources for mobility traces and methods for analyzing these data, used in this thesis. As datasets may vary in size, sampling frequency and even accuracy, robust statistical methods are required to model the studied behavior. Finally, these models can be used to either reproduce different aspects of mobility, or yield valuable insights about their underlying processes.

Next, we present results of our studies on how mobile devices can be used to understand mobility and how their use is correlated with network usage, mobile applications, and human contacts.

3 Understanding mobility through mobile devices

In this chapter, we present results and discussions aimed at better understanding human mobility through trace analysis. For that, a series of models are devised, allowing us to explain nuanced aspects characteristics of how people move as well as direct consequences of their mobility. Given the popularization of mobile smart devices, such as laptops and smartphones, a growing amount and variety of behavioral data can be obtained, which constitutes directly or indirectly the sources of all traces used in this chapter. We study human mobility in four parts. Under (**§ 3.2**) we discuss the link between mobility and phone utilization, namely network traffic usage and online video games playing with a strong mobility aspect, showing strong relationships between physical and virtual world of mobile users, while we also model the predictability of human movements, enabling us to build a theoretical foundation as well as limits for how well an ideal forecast model should be expected to achieve. Furthermore, under (**§ 3.3**), using a variety of datasets, we study how online mobile games may influence mobility. On (**§ 3.4**), based on the aforementioned understanding of mobility, we look into byproducts of this mobility, namely visit and contact duration patterns that emerge as a consequence of human mobility. Finally, under (**§ 3.5**), we study how to improve the prediction of the dynamics of an epidemic spreading. In this chapter, we answer **RQ1**, on the relationship between human mobility and smartphone usage.

To better understand the existing foundational work in these areas, we present next a summary of the state of the art, including work that is relevant to the results and discussions presented later in this chapter.

3.1 State of the art

We discuss a series of articles, organized by topics associated to results we present in the next sections.

3.1.1 Human Mobility

Studying the spatial-temporal aspects of human mobility, the seminal work by Gonzalez *et al.* [11] and Song *et al.* [36] reveal the various aspects of regularity and limits for forecasting urban movements using mobility traces based of cellular networks. The pervasiveness of smartphones and their growing list of sensors enabled researchers to study various aspects of *human mobility* in the last 20 years. Random models for movements,

such as Random Walk, were replaced by Lévy-flight models [67, 11] in which the distributions of hops (or flights) follow a power-law distribution (see Chapter 2). Using data sets with higher resolution, such as GPS based studies, these observations have been more recently revisited. These studies conclude that the distribution of displacements follow a log-normal distribution [44, 45] in urban scenarios while exponential in intra-urban trips [68].

Human mobility has also been modeled around social interactions [17, 53], natural disasters [69], and income [70]. Additionally, information dissemination is a fundamental aspect of mobility that has been largely studied, either for opportunistic data forwarding [71] or contagious disease spread modelling [72, 73]. The seminal work by Hui *et al.* [74] revealed long-tailed distributions in *inter-contact time* (time interval between consecutive contacts of any pair of devices) instead of exponential distribution and its implications on opportunistic forwarding systems using a data set collected during a scientific conference.

3.1.2 Device Variation

Usage and traffic patterns of different device types have been studied from various perspectives (*e.g.*, [75, 76, 77, 78, 79, 80]). For example, in [81], authors use packet-level traces from 10 phones and application-level monitoring from 33 Android devices to analyze smartphone traffic. Furthermore, the study by Das *et al.* [82] analyzes 32k users on campus, and focuses on multi-device usage while noting differences between laptops and smartphones in packets, content, and time of usage. Their work targets device usage patterns and security. They observe that the usage of laptops and tablets is interchangeable and that among multiple devices, their usage is additive instead of being shared among devices.

3.1.3 Online activity with effects on mobility

Studying a large set of WLAN traces, Cao *et al.* [83] revealed surprising patterns on increases of long-term mobility entropy by different demographics, such as age and the academic majors of students. The work by Moghaddam *et al.* [84] characterized web domain access by users and their respective locations, drawing correlations between NetFlow and DHCP traces from a large university, resulting in a realistic set of scenarios for simulation.

Location-based gaming has steadily emerged as a popular pastime on smartphones, and has become a potentially effective way at promoting physical activity [85, 86, 87]. From a scientific standpoint, the unique and most interesting aspect of these games is how they encourage and promote movement, which can improve physical and mental health [88, 89], and be comparable to a health or a fitness app [85, 86, 87, 90, 91]. More generally, location-based games are examples of a broader class of smartphone applications that attempt to promote physical activity – either directly through recommendations or indirectly through objectives that are linked with physical locations [85, 86, 87, 92]. Other examples of applications in this category include varied location-based

services [53], online location-based social networks [93] and smartphone and wearable applications for physical activity [94].

3.1.4 Mobility Predictability

The seminal work by Song *et al.* [26], utilizing cellular network data, established an approach towards understanding and measuring the predictability of human mobility patterns, with their equally important contribution with respect to the data-driven analysis of large mobile populations, and their efforts in devising a framework to study the theoretical limits of predictability. The methods introduced in their framework are founded on information theory and have since been extensively applied in the area of mobility modeling and prediction. Later studies that built on the work by Song *et al.* [26] addressed either the specifics of the prediction problem (*e.g.*, different formulations of the individual's change of location [95]) or the shortcomings of the original approach that relied on coarse spatio-temporal granularity. Cao *et al.* [83] used Wireless LAN (WLAN) traces from a university campus network and reported multi-modal entropy distributions which can be partially explained by the demographics of the population (*i.e.*, age, gender, major of studies). Other entropy based studies include vehicular mobility [96, 97, 98], online social behavior [99, 100], complex systems [101], cellular network traffic [102] and public transport utilization [103].

3.1.5 Contact Duration

The study by Chaintreau *et al.* [10] includes 8 different data sets on human contacts. Other similar studies include fine-grained measurements, such as schools [104], conferences and museums [105]. The work by Sun *et al.* [25] studies contacts using a metropolitan scale data, but limited to public transport. Common limitations of such datasets include the lack of accurate location information, or being limited to small spaces (*e.g.*, conference venues and university).

While short *inter-contact times* are associated with lower latency in opportunistic networks, large *contact duration* can be seen as high throughput [106]. Regardless of their importance, most recent studies have focused on the former, mainly as recent advancements in wireless network technologies brought a larger bandwidths to mobile devices, even though data exchange capacity grows as contact duration gets longer. When modelling the spread of infectious diseases, however, *contact duration* is a key aspect [107, 104]. *Contact duration* allows the study of how epidemics spread through a temporal network, in which edges between nodes evolve over time [108]. While such studies often better describe the dynamics of diseases outbreaks and their prevention, little is still known about how mobility and contacts are related.

3.1.6 Graph Representations and Epidemic Forecasting

To efficiently study how information spreads among individuals, a network approach is often used for its low complexity and high effectiveness. The sharing of files in a network

or the spreading of an infectious disease can be modeled and studied using varying numbers of individuals, or nodes, connected through links which may represent a connection between two computers or a prolonged contact between two persons. However, to build predictive models of these networks requires a change in how this inferred network is represented, where one commonly used approach is node embedding [109].

In recent studies, considering weak ties was helpful in different sociological contexts, such as the influence of indirect contacts on decision-making [110], or the dismantling of organized crime [111]. In a more graph-theoretical approach, researchers explored different methods of clustering people by their roles, relying on the fact that structurally equivalent nodes fulfill the same role in society [112].

While most node embedding methods focus on homophily, some have been developed to preserve only structural equivalence. The method proposed by Wang *et al.* [113] has a precise mathematical approach, while `struc2vec` [114] is based on random walks similarly to `node2vec`. Another popular embedding method uses the recursive nature of structural equivalence to create an embedding, as two nodes are considered structurally equivalent if their neighbors are structurally equivalent [115].

While reviewing node embedding methods, Junchen *et al.* confirmed that `node2vec` embeddings perform best against other methods at preserving local structure [116], whereas Schliski *et al.* included notes of caution by testing `node2vec` with different hyperparameter settings and concluded that `node2vec` does not preserve structural equivalence well, even with outward oriented hyperparameters [117].

3.2 Online activity and mobility

In this section we present results on how human mobility and online activity are intertwined. These analyses highlight the importance of studying both aspects combined, enabling insights that are not possible when each part is studied in isolation. We begin by looking at how properties of human mobility and network traffic correlate in a university campus Wi-Fi network. This first analysis allows us to establish a relationship between how mobile users consume network resources, especially regarding the type of mobile device they are using. This part is then followed by results on how mobile online games with a strong mobility component influence the daily movement of players. These latter results demonstrate how exogenous factors, such as the usage of a mobile device functionality, may explain deviations in the regularity and predictability of human mobility, such as those observed in the former results of this section.

This study includes analysis of both mobility and traffic for a large set of mobile devices in a university campus environment. Furthermore, using the Organizationally Unique Identifier (OUI) (*i.e.*, Media Access Control (MAC) manufacturer prefix) as well as traffic pattern from devices, we further classify them into two larger groups, namely smartphones and laptops. This classification allows us a unique view of how mobility and traffic differs between these two classes.

3.2.1 Datasets

The traces used for this study were chosen to represent how smartphones and laptops move in space and their corresponding network traffic information, where any UID was the same in both sets. For this, we used two datasets: (**DHCP**) containing Wi-Fi Access Point (AP) logs, and **Netflow** records. For details on the sizes of these two sets, see Table 3.1, and for a sample of records from each set, see Table 3.2.

3.2.1.1 DHCP

These association and authentication logs were collect from 1760 APs, distributed in 138 buildings for a total of 479 days (≈1.3 years) at the University of Florida campus. This set includes over 550 million events from 316,000 devices in the years 2011 and 2012. Each record includes the device's MAC address (*i.e.*, its UID), assigned IP address, the associated AP and timestamp of the event. This information allows us to approximate the location of a device by the coordinates of a building where an associated AP is found. We further validated these associated locations with a crowdsourced service for APs, *wigle.net*. We were able to match a total of 130 routers, distributed across 58 different buildings, and all APs were within 200 meters (or less) of their original mapped location.

3.2.1.2 NetFlow

A total of 76 billion *flows* (*i.e.*, Netflow records) were sampled from the same university network over 25 days in April 2012. Each *flow* includes origin/destination pairs of Internet Protocol (IP) addresses and ports as well as transport protocol.

The *NetFlow* logs of a device are matched to their corresponding location records (*DHCP*) using its IP address, commonly present on both records. We refer to this final database as *CORE*, which also contains location, as described above, and web domain information using reverse DNS.

Table 3.1: Summary of datasets. mil=million bil=billion.

	Records		Traffic Volume		Devs	
	DHCP (mil)	CORE (bil)	TCP (TB)	UDP (TB)	DHCP (K)	CORE (K)
Smartphones	412.0	2.13	56.18	4.50	186.0	50.3
Laptops	101.0	4.20	73.85	12.90	93.2	27.1
Total	557.5	6.53	134.39	17.61	316.0	80.0

Note that we were unable to classify all devices, while for others we identified them as printers or other equipment not relevant to this study, therefore those were removed.

3.2.2 Device classification

The classification of devices according to their type is done is several steps, which we explain next. The first important observation is that the first three octets of a MAC address uniquely identify its manufacturer, *i.e.*, its OUI. Even though some manufacturers may produce multiple device types, an OUI is typically assigned to a single type. Next, we conducted a survey to identify their device types, during which we collected the original MAC of 30 devices from users who agreed to share that information. Finally, using the OUI and survey data we were able to identify 46% of the total devices, with 90,000 laptops and 56,000 smartphones.

To classify a larger number of the remaining devices, we observed that over 3,000 devices had `admob.com` in their *CORE* records, out of which 92% were smartphones previously classified. The `admob.com` domain was used by Google to provide in-app ads to both Android and iPhones, therefore being an exclusive service targeted to mobile handheld devices at the time. This observation allowed us to further classify over 270,000 devices, accounting for 86% and 97% of total devices in the *DHCP* and *CORE* datasets, respectively.

Table 3.2: *DHCP* (top) and *NetFlow* (bottom) sample data.

Device IP	Device MAC	AP name	AP MAC	Lease begin time	Lease end time
10.131.97.9	be:ef:ca:fe:15:07	b4r14-win-1	00:1d:ff:8f:bc:aa	133323	133353

Start time	Finish time	Duration	Src IP	nnnnnnDst IP	Protocol	Src port	Dst port	Packet count	Flow size
133433.912	133933.576	1.664	9.54.37.7	10.15.25.126	TCP	80	60482	157	217708

3.2.3 Mobility trace analysis

We now discuss our observations about the differences in the mobility of smartphones and laptops.

3.2.3.1 Session start probability

The start of a session happens when a device first associates to an AP, and captures the expected activity level of devices at a given location. In our analysis, the start times of sessions align with the periodic beginning of classes, especially in *Academic* buildings. In such locations, activity drops strongly for laptops at 5pm, whereas smartphones keep up higher levels of activity until 8pm. In buildings of type *Social* and *Library*, activity remains higher late into the evening, with a smoother decay in levels as devices exit the network. Similar observations, however, do not hold during weekends when schedules are no longer influenced by lectures.

3.2.3.2 Radius of gyration and other spatial metrics

Studying the radius of gyration for the devices in our dataset, we make the following observations: (**1**) After a transient period that lasts one week, the radius of gyration

stabilizes over several months for both device types. (**2**) When comparing weekdays and weekends, laptops see a significant reduction in mobility at the end of the week, whereas smartphones only sees mild changes, which could be explained by differences in how each device is used. (**3**) In spite of covering a large area (8.1km^2), the expected distances covered by half of smartphones and laptops are relatively small, with 295 meters and 172 meters, respectively. Other spatial metrics were also evaluated, such as *diameter*, the *maximum distance* between two consecutive records, and the total sum of trips done by a device which have shown a similar behavior to the radius of gyration.

3.2.3.3 Preferred locations

To better understand the preferences of a user when visiting different places, we study the number of unique places visited as well as the time spent at the most visited location. This information allows us to model not only the area covered, but the places a user could be found, where there could be opportunities for connecting to the wireless network. From our analysis, we observe laptops being used for longer periods of time, although with a similar median to smartphones of 2 hours and 40 minutes. This similarity coupled with a lower number of total visited locations by laptops suggests that these devices are preferably used when users stay put for longer periods of time.

3.2.3.4 Exploration

Interesting differences are observed between phones and laptops when we study their users' tendency to explore new places or return to previously visited ones. We observe that after a transient period of 7 days, the rate of visits to new places is similar for both device types, with a significant change after 120 and 240 days, which aligns with typical 3-months blocks of each university term (see Figure 3.1). Furthermore, we study the probability of finding a device back at its *L-th* most visited location (*i.e.*, building), where we observe a Zipf distribution [11] with coefficients $L^{-1.36}$ for laptops and $L^{-1.16}$ for phones. These results support those found by Gonzalez *et al.* [11] based on CDR records. These coefficient differences capture the more exploratory nature of smartphones, which are expected to be used in more varied and new locations. Conversely, the use of laptops tends to be more strongly associated with fewer, but more frequent places.

3.2.4 Mobility prediction

For this analysis, we compared results of two deep learning methods and two entropy rate estimators for both device types. The deep learning methods used were Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) [118] for their proven record in predicting speech and text sequences. Additionally, we compare entropy rate values of the LZ and BWT estimators (see § 2.2.4).

To use the aforementioned methods, we first construct a discrete time series from the sequence of visits of a device.

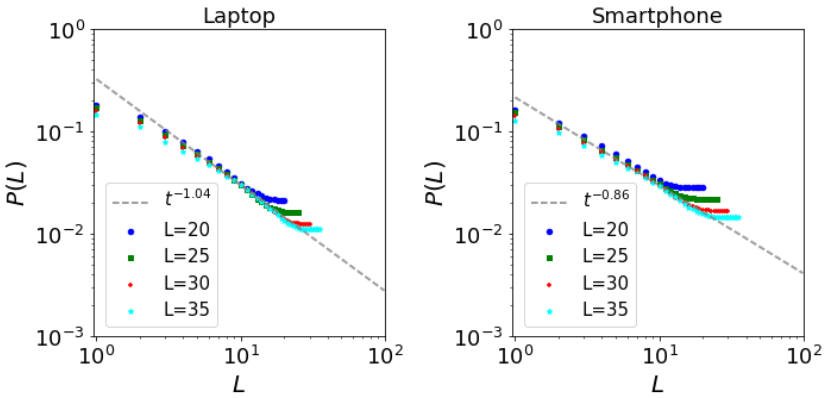


Figure 3.1: Zipf's plot on L visited access points.

3.2.4.1 Discrete Time Series

Given a set a of timely ordered events $X = \{x_t : t = 1, \cdots, n\}$, where x_t is the realization of X at time t for $t \in T$, we say that a timeseries is *discrete* if T are measurements taken at successive times spaced at uniform intervals w, also referred to as sampling rate (defining the temporal granularity). In this work, each measurement corresponds to a network AP event (*e.g.*, association), which records a location at a given timestamp. Note that, for the studied setting (*i.e.*, university campus Wi-Fi network), the choice of w allows noise filtering (*e.g.*, mitigating ping-pong associations between neighboring APs) as well as precising timing of a location change. Similarly, spatial resolution can mitigate noise but also define the precision of a device's location. The choice of these parameters is influenced by the settings under which the dataset was collected along with the application being considered in the study.

3.2.4.2 Accuracy and predictability

Table 3.3 shows the results of the prediction algorithms as well as the theoretical predictability from the studied entropy estimators. The table summarizes the median values for different device types, and spatial and temporal granularities. As expected, our results show that *laptops* are more predictable than *phones*, mainly for the higher levels of mobility of the latter as previously discussed. Feeding the prediction algorithms with longer sequences do not hinder their performance significantly, and in all cases, the theoretical bounds given by LZ and BWT were not yet matched by the neural networks, suggesting that better approaches are still possible even without considering a wider set of features (*e.g.*, schedule or weather forecast).

3.2.5 Network traffic

Based upon our analysis we note that network traffic is significantly different between phones and laptops. Mainly, we observe that, on average: (1) laptops consume 2.7x the traffic volume of phones, in spite of phones' flows being 2x larger, (2) phone packet

Table 3.3: Median Accuracy for *phones* vs *laptops* (Diff is *laptops* - *phones*).

		AP, 1h			Building, 1h			AP, 15min			Building, 15min		
Seq Len	Predictor	Phones	Laptops	Diff	Phones	Laptops	Diff	Phones	Laptops	Diff	Phones	Laptops	Diff
5	LSTM	21.62	25.00	+3.38	35.03	50.00	+14.97	40.00	44.56	+4.56	52.44	65.56	+13.12
	CNN	16.45	24.27	+7.82	34.94	50.00	+15.06	50.00	59.80	+9.80	64.60	76.94	+12.34
10	MC	17.98	25.6	+7.62	36.72	50.28	+13.56	52.25	61.97	+9.72	68.00	82.25	+14.25
	LSTM	20.83	26.31	+5.48	37.50	50.66	+13.16	31.14	44.62	+13.48	45.38	64.56	+19.18
	CNN	18.06	22.62	+4.56	36.20	52.03	+15.83	49.20	58.80	+9.60	64.56	74.00	+9.44
20	LSTM	21.22	24.19	+2.97	36.12	50.78	+14.66	29.17	41.00	+11.83	43.62	61.47	+17.85
	CNN	18.44	23.60	+5.16	35.28	50.00	+14.72	37.84	48.12	+10.28	50.00	65.00	+15.00
40	LSTM	19.67	24.33	+4.66	32.62	52.03	+19.41	23.30	39.40	+16.10	33.97	59.03	+25.06
	CNN	18.75	23.97	+5.22	35.25	52.50	+17.25	27.62	44.70	+17.08	41.25	62.10	+20.85
	LZ	46.90	52.60	+5.70	58.78	66.40	+7.62	72.70	76.06	+3.36	79.60	79.10	-0.50
	BWT	66.44	69.44	+3.00	73.70	79.90	+6.20	83.30	88.06	+4.76	88.60	92.20	+3.60

sizes are 50% larger than those used by laptops, (3) both device types show a similar active flow runtime, despite the aforementioned differences in flow size and total traffic volume, (4) both device types show comparable inter-arrival time for packets, however, with a higher deviation for phones, and (5) no major differences between protocols being used. We also note that, unsurprisingly, there were no significant differences between week periods. Furthermore, the observed differences between phones and laptops are likely due to any mobile Operating System (OS) optimizing resources for better battery consumption.

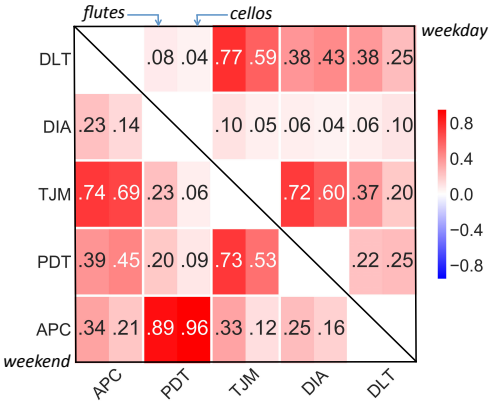
3.2.6 Mobility and traffic combined

We now study how mobility and network traffic may be interconnected, highlighting the importance of considering both aspects when analyzing the behavior of mobile users. For this comparison, we compute the correlation per device between selected mobility and network traffic features. This selection is done using Correlation Feature Selection (CFS) [119], which excludes cross-correlated features but keeps those strongly correlated to the target classes. For a table of abbreviations used to present the results along with correlation results between features of the same class (*i.e.*, mobility *or* traffic), see Figure 3.2.

For the 8 studied mobility features, CFS yielded 5 features as relevant for the analysis against traffic features. Similarly, out of the 19 network traffic metrics, CFS reduced them to 11 given their relevance to the comparison with mobility features. It is interesting to note that the mobility correlations suggest that users spend most of their weekends at their preferred buildings, such as libraries, and devices which spend more time online do not necessarily consume more traffic with no difference between times of the week.

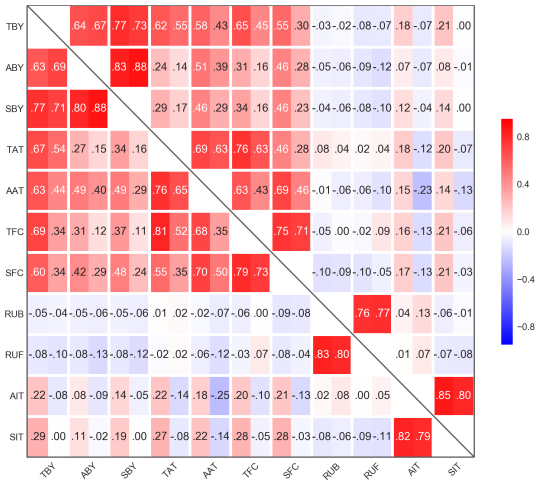
3.2.6.1 Integrated analysis

By combining these two sets of features, *i.e.*, mobility and traffic, we observe that, as the number of locations visited increases, so does the expected time devices remain active. In contrast, an increase in that same mobility metric correlates negatively with



Abbr.	Description
APC	AP Count (unique)
PDT	Preferred building Δt
TJM	Total (sum) jumps
DIA	Diameter of mobility
DLT	Delta time (time of network association)

(a) Mobility



Abbr.	Description
TBY	Total flow bytes
ABY	Avg. flow bytes
SBY	Std. flow bytes
TAT	Total active time
AAT	Avg. active time
TFC	Total flow count
SFC	Std. flow counts
RUB	UDP bytes / total bytes
RUF	UDP flows / total flows
AIT	Avg. IAT
SIT	Std. IAT

(b) Traffic

Figure 3.2: Correlation for *mobility* (a), and *traffic* (b).

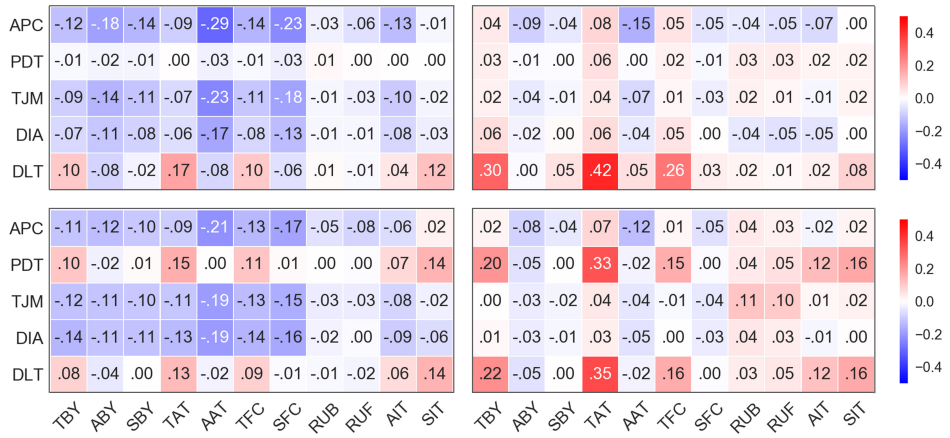


Figure 3.3: Correlation mobility-traffic. Weekdays (top) and weekends (bottom), for phones (left) and laptops (right).

the total (and standard deviation of) flow counts for both device types. Furthermore, for both device types, we observe no significant change in traffic consumption with more frequent movements, which suggests that mobile users do not necessarily browse more or less for being more mobile. Finally, using various machine learning approaches we observe the potential for an integrated mobility-traffic model by building various models to classify devices according to their type: (1) using *supervised learning* (*i.e.*, Support Vector Machine (SVM)) on the device label data, we observe an increase from 65% and 79% up to 81% when considering mobility, traffic and them both combined, respectively, and up to 86% when considering the time of the week (*i.e.*, weekdays/weekends), (2) similarly, using *unsupervised learning* (*i.e.*, k-Means) prediction accuracy goes from 60% and 81.2% up to 81.5%. These results suggest that future predictive models for mobile device usage could benefit, at least in some cases, from considering *mobility and traffic* features in combination. That is, these two aspects of human behavior may be associated and therefore, may influence each other at certain times. It is important to ww that the results may be subject to bias due to the method of data collection, based on Wi-Fi associations and which are subject to how often the devices were used.

These results highlight the importance of studying human mobility in combination with network traffic as the two set of activities are interconnected. That is, even though even phenomenon may be studied independently, when both are considered at the same time, we are able to better model and understand these aspects of human behavior. Next, we look at how human mobility and online games are related.

3.3 Online games and mobility

We now turn to the study of how human mobility and online games could be interconnected. As in the previous section, we have shown how network traffic and mobility can produce better behavioral models, we now look at how one type of application may have influenced how its users move. For this, we focus on the mobile game Pokémon

GO for its strong mobility component, in which gamers are incentivized to move around urban spaces to collect items in the game. This study was done using a combination of datasets, including mobility data from an OSN and individual traces from a monitoring app, capable of tracking phone usage and mobility of subjects.

The main game studied, Pokémon GO, was released in July 2016 and quickly became one of the top downloaded and used apps for both iOS and Android devices. Users could interact with a virtual reality world by moving around their real urban space, collecting items and performing activities at selected physical locations only. This lead gamers to use their phones at various locations, and as reported by previous studies, increase the number of steps taken in a day [85]. This reported increase in daily movements raised the question of how gamers actually moved with respect to the distribution of distances traveled. As human movements tend to follow a long-tailed distribution, often associated to Lévy-flight, our study focused on investigating whether or not these distributions remained unchanged or how they could have been modified by the game. To isolate the effects of the game, we compared various aspects of human mobility measured before, during and after players used the game, as well as compared similar effects for gamers of another app and non-gamers.

3.3.1 Datasets

We study the mobility changes using two main datasets: (1) with data collected through Carat, an energy monitoring app, and (2) with data from Twitter, a popular OSN. The first set includes mobility and phone usage data from January 2016 to March 2018, while the second set includes mobility data from January 2016 to June 2017, covering several months before and after the release of Pokémon GO in July 2016.

3.3.1.1 Carat

This first dataset was collected from users of the Carat application[1] [9]. This software logs multiple aspects of a smartphone with minimal interference, including current settings, battery and connectivity state, list of currently installed and running apps, and distance traveled since the last update. Each record is captured with every 1% change in battery level from each device, and is stored with a UID and timestamp. Finally, the Carat application uses this information to recommend its users tailored suggestions to increase their devices' battery lives [120]. It is important to note that Carat does not log any data while running on the background, but relies on the device's OS to register battery change events. There are, however, some scenarios which may lead Carat to no be able to record any events, leading to a sparser dataset. These scenarios include when the phone is on battery save mode (or deep sleep), when the OS evicts Carat from memory to save up resources, or when Carat is closed by the user.

Carat was first released in 2012 and has been installed in over one million phones in several countries, for which we use data from January 2016 to March 2018. For this study, we consider only Android users as during the desired period, iOS no longer reported the

[1] `http://carat.cs.helsinki.fi/`

list of running applications. This dataset contains 173.6 million entries, from 74,000 phones. To analyze the effect online games may have on human mobility through Carat, we contrast the same measurable effects between two gaming apps, namely Pokémon GO – a location-based game – and Clash Royale[2] – a game without a physical world component. For the former, Carat contained 3,996 users while for the latter 1,323 users. For a summary of the number of users per country (top-10 only) on the Carat dataset, see Table 3.4.

It is important to note that Carat computes the distance between two consecutive records in a privacy preserving manner. For that, at every new event Carat samples the current location of a phone and computes its distance from the previous record. Finally, only this computed distance is permanently stored and transmitted to the back-end while old locations are destroyed. Furthermore, the aforementioned locations are sampled from the Location Manager Application Programming Interface (API) of Android, which only provides coarse-grained locations with an inaccuracy of up to 2 km.

Although the data collected has a low spatial resolution, it still provides relevant data points that are useful in understanding the overall trends and patterns of the phenomenon being studied. This is because the data points can still capture the larger-scale dynamics and relationships between variables, even if they may not capture smaller-scale variations.

3.3.1.2 Twitter

This second dataset was collected from users of Twitter, a microblog OSN in which posts (publications, or *tweets*) may also contain location information of the devices being used to tweet. We analyze 8.7 million posts containing location information, from more than 21,500 Twitter users located in 15 countries. This larger set of users and diverse number of countries was chosen to weaken the influence of potential location biases. These records were obtained through Twitter's open API[3], from where only tweets containing geographical information were kept, accounting for 17.4% of the total. Before downloading each post, the search criteria was the following: (1) the name of a large metropolitan area, *e.g.*, Bangkok, Thailand, and (2) a period within the time of the study. To classify users as *gamers*, we look for any variation of the string *#pokemongo* in their tweets (*e.g.*, #pokémongo, #PokemonGO), yielding over 8,900 *gamers*. To validate this step, we manually inspected 1% of randomly sampled tweets from *gamers*. In this validation step we observed that 90% of gamers' tweets with *#pokemongo* had some content related to the game, *e.g.*, screenshots or comments about the game. Additionally, we filter out possible bot accounts with the *Botometer* [121], removing 3.1% of accounts classified as non-humans. For a list of total *gamers* per city, see Table 3.4. Both *gamers* and *non-gamers* showed similar statistical properties regarding their tweets behavior. On average, the number of tweets was 390 and 351, with medians 200 and 159 for *gamers* and *non-gamers* respectively, and their distributions were statistically similar (*i.e.*, *p-value* < 0.001). Similarly, the inter-arrival-time for tweets was statistically similar between the

[2]https://clashroyale.com
[3]https://developer.twitter.com/

two groups, with 57.8 hours and 58.1 hours on average and 7.68 hours and 8.13 hours median for *gamers* and *non-gamers*, respectively (p-*value* < 0.001). We note that in all studied cities, both groups presented a similar spatial distribution of points, with a strong presence in urban areas.

Table 3.4: Left: Number of gamers on Twitter. Right: Number of gamers on Pokémon GO (PG) and Clash Royale (CR).

Twitter

City (code), Country	N.
São Paulo (SPO), Brazil	924
Jakarta (JKT), Indonesia	911
London (LON), UK	853
Singapore (SIN), Singapore	709
Santiago (SCL), Chile	661
Tokyo (TKY), Japan	631
Bangkok (BKK), Thailand	599
San Francisco (SFO), USA	597
New York (NYC), USA	564
Toronto (TOR), Canada	447
Paris (PAR), France	373
Seattle (SEA), USA	348
Boston (BOS), USA	279
Sydney (SYD), Australia	268
Hong Kong (HKG), China	263
Barcelona (BCN), Spain	247
Moscow (MOW), Russia	143
Helsinki (HEL), Finland	92

Carat

Country (code)	PG	CR
USA (us)	780	134
Finland (fi)	746	175
Germany (de)	495	79
UK (gb)	153	20
Canada (ca)	149	20
India (in)	122	137
Japan (jp)	113	6
Spain (es)	102	58
Italy (it)	78	43
Netherlands (nl)	50	9

3.3.1.3 Google Trends

In addition to the utilization of Twitter and Carat datasets for the analysis of mobility patterns, we incorporated data from Google Trends to study the temporal dynamics of mobile game popularity. This approach was chosen for the insights provided by this service, which generates a trendiness score (G) for a specified search term across various time frames and geographic regions. By incorporating the trendiness scores, we gain a temporal understanding on the interest in the mobile games, helping select which periods changes could have been strongest in the users' mobility.

The heterogeneity and size of datasets used ensure statistically sound observations as well as insights from different perspectives on the influence of location-based games on human mobility. Each of the two main datasets provides a unique yet complementary viewpoint of how mobility could have changed during the period of peak popularity of the game.

3.3.2 Analysis

We now focus on the series of pre-processing steps and analysis taken to study how location-based games may alter human mobility.

3.3.2.1 Spatial noise filtering – Twitter

We begin by eliminating geographical points of a user's trajectory that fall outside a their own city. This filtering step is required for two main reasons: (a) exclude visits to other cities and countries, while focusing in the main area visited by a user, and (b) to limit the range of motion covered by any studied trip as those influenced by the game typically cover up to 10 km (or 6.2 miles) [122, 123]. Note that with both (a) and (b) we limit our study to commonly visited areas for each user, we therefore cluster Twitter traces as being either *local* or *away* with respect to the city assigned to each subject. This assignment is done based on the city from which a user was initially found at inside Twitter. To achieve that, we classify a city C in four steps, where we let C_S denote all geographical points at or near C from all available Twitter users: (1) we combine nearby points in C_S using DBSCAN [124] with $\epsilon = 2$ km (*i.e.*, maximum allowed distance between any two points in the same cluster), (2) we compute the center of mass C_{cm} of the cluster with the highest number of points from step (1), (3) we compute the distances between every point in C_S and C_{cm}, namely $d_{s,m}$, and (4) we cluster the log transformation of all $d_{s,m}$ using KMeans with $k = 2$ (*i.e.*, the number of expected clusters to be found). Both of these clustering algorithms were chosen for their robust unsupervised approach in finding classes on geospatial data. While DBSCAN groups points based on a maximum distance threshold, KMeans classifies points among a predefined number of clusters. In all 18 studied cities the distribution of the resulting distances to C_{cm} showed a clear segmentation between *local* and *away* points around 100 km (62 miles). A possible explanation to this clear separation are the distances people commute regularly, regardless of location. Finally, for any city C, we consider only *local* tweets from users who had a minimum of 25% of their posts at C.

3.3.2.2 Place Extraction – Twitter

To study users mobility, we group consecutive hops that are part of a single trajectory by first identifying relevant locations in their data [125]. A series of points define a *stop* when: (1) they contain no displacements, (2) the time between samples is lower than τ, and (3) the considered sequence has a minimum duration τ. Similarly, we define a *movement* using (2) while being bounded by *stops* (*i.e.*, preceded and followed by a *stop*) within a maximum interval τ. We set $\tau = 15$ minutes to capture short stays while discarding very short stops (*e.g.*, at traffic lights), further reducing the uncertainty about when a movement happened.

3.3.2.3 Temporal Analysis – Multiple sources

To choose the best period for our study based on the popularity of the game, we consider four observations: ($\boldsymbol{I}$) the number of installations from Carat, ($\boldsymbol{S}$) the number of gaming sessions from Carat, ($\boldsymbol{N}$) the number of tweets containing `#pokemongo`, and ($\boldsymbol{G}$) the Google trend index for Pokémon GO, which can all be observed in Figure 3.4. This allows us to define three main periods of interest: *before* the game (April-June/2016), *during* (July-September/2016) and *after* (October-December/2016), and therefore consider only *gamers* with records in all three periods. It is noteworthy that the expected time between the first and last tweet about the game was significantly shorter than the expected time people played as seen on Carat, with 59.2 days vs. 99 days, respectively. In spite of these differences, the powerlaw exponent for the distribution of interval interacting with the game on Twitter and Carat was similar, with α equals 1.285 and 1.305, respectively. We conjecture that players likely stopped tweeting about the game after using it for a while.

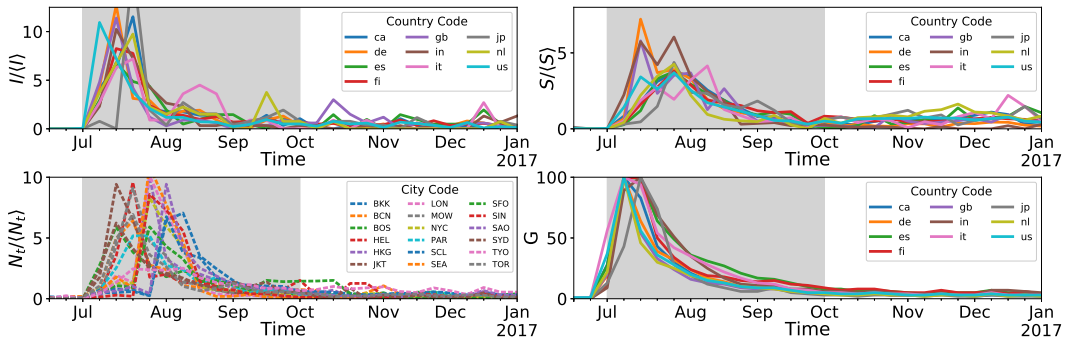


Figure 3.4: Pokémon GO new installations (I), game sessions (S), number of tweets (N_t), and Google Trend index (G). Where the first three are normalized by their average ($\langle\bullet\rangle$).

3.3.2.4 Number of visited locations (φ) – Twitter

To reduce spatial uncertainties on tweets, we combine them into discrete points formed by a squared mesh with 250 meters sides per cell. We observe a stretched exponential (*i.e.*, Weibull) distribution for the number of visited locations, with a stretching exponent β close to 1, which allows us to approximate this distribution to an *exponential*. With that, we can write that the expected number of visited locations (φ) by a user after time t is $\varphi(t) = 1/\lambda(t)$. We will use this simplification to study deviations in the behavior of *gamers* compared to *non-gamers*. The fit parameters for *gamers* and *non-gamers* were $\lambda = 0.0226$ and $\beta = 0.946$, and $\lambda = 0.0193$ and $\beta = 0.916$, respectively, with a notable higher visitation average ($1/\lambda$) for *non-gamers*.

3.3.2.5 Gaming session (S) – Carat

Using data from Carat, we group consecutive records containing a game running (*i.e.*, Clash Royale or Pokémon GO) into a single gaming session, as long as the time between each record is no longer than 5 minutes. Given the sparsity of the Carat dataset, we only consider gaming sessions that are at least 5 minutes long. Interestingly, the distribution of session duration of both games for various countries is similar, highlighting the nature of a fundamental behavior between *gamers* across different areas and cultures. The distributions of $\Pr(S)$ for both games, in various countries, are depicted in Figure 3.5 and Figure 3.6.

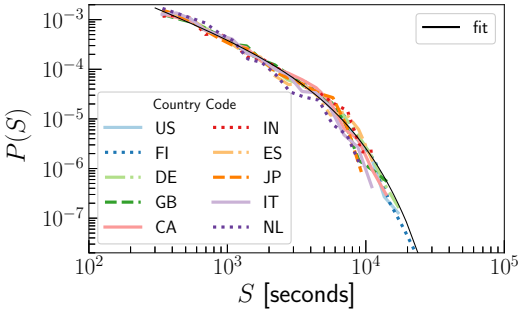


Figure 3.5: Game sessions duration **Pokémon GO**.

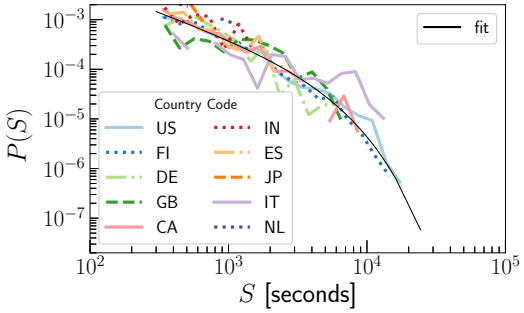


Figure 3.6: Game sessions duration **Clash Royale**.

3.3.2.6 Distance traveled between consecutive records (Δr) – Twitter

We define the distance between consecutive tweets (Δr) as the straight line between those two points, discarding improbable velocities for urban environments above 120 km/h (75 miles/h). For fixed time intervals (ΔT_o), the distribution of distances traveled remains stable, showing a robust behavior to different sampling sizes, depicted in Figure 3.7. Furthermore in the same Figure, we observe a multi-modal distribution of Δr of a truncated powerlaw (*fit 1*) and an exponential (*fit 2*), split at an inflection point around 30 km (18.6 miles). The parameters for these distributions also similar for both *gamers* and *non-gamers*, as shown in Table 3.5. While *fit 1* suggests Δr and its probabilities are proportional up to a cut-off, *fit 2* entails a fast drop in the probabilities of Δr in which higher values can no longer be observed.

Table 3.5: Fit parameters for Figure 3.7.

Twitter Users	Truncated Power Law (fit 1)	Exponential (fit 2)
Gamers	$\alpha = 0.279$, $\lambda = 0.089$	$\lambda = 0.036$
Non-gamers	$\alpha = 0.329$, $\lambda = 0.083$	$\lambda = 0.036$

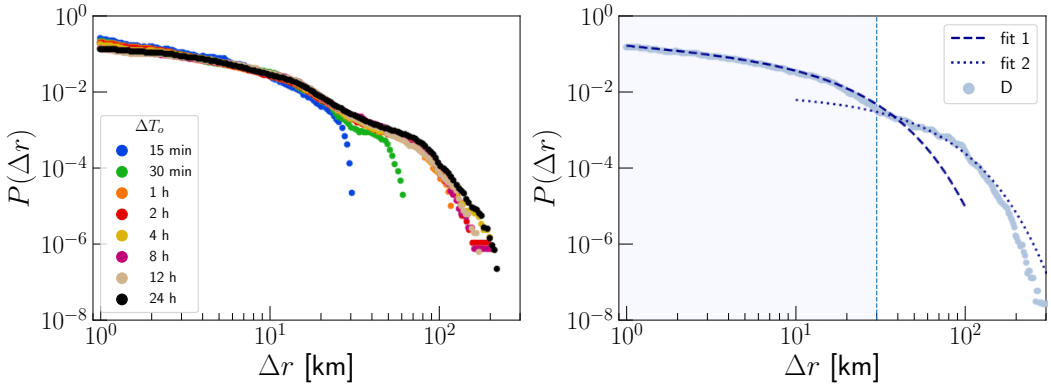


Figure 3.7: Distribution of displacements for the Twitter dataset. (Left) $P(\Delta r)$ over an interval ΔT_o. (right) Multi-modal fit, truncated powerlaw (fit 1) and exponential (fit 2).

3.3.2.7 Δr – Carat

As no geographical coordinates were available with Carat, we limit displacements in this set to 100 km (62 miles). Interestingly, using Carat we also observe a multi-modal distribution for Δr as depicted in Figure 3.8, however, with different parameters: *fit 1* $\alpha_{\text{twitter}} = 0.329$, $\alpha_{\text{carat}} = 1.469$), $\lambda_{\text{twitter}} = 0.08386$, $\lambda_{\text{carat}} = 0.08346$); *fit 2* $\lambda_{\text{twitter}} = 0.03589$, $\lambda_{\text{carat}} = 0.04505$. The differences in the observed parameters are likely due to how each set samples displacements. That is, while Twitter users tend to post about the game at a final destination, Carat might sample a trip while it is still happening, leading to more granular samples.

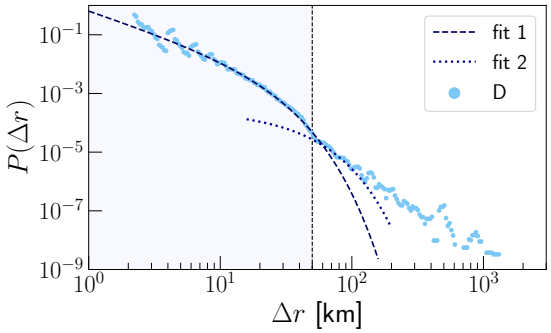


Figure 3.8: Distribution of displacements for the Carat dataset. $P(\Delta r)$ with a multi-modal fit by a truncated powerlaw (*fit 1*) and an exponential (*fit 2*), split at 50 km.

3.3.2.8 App battery performance may hinder usability and mediate mobility change – Carat

We correlate changes in mobility and different versions of Pokémon GO, along with insights available in the game's changelogs. Using the Carat data set, we first look at

the adoption rate of each version during the initial four months of its first release. We note that, after battery issues were addressed (v0.31 and v0.33), players who started with these versions showed a significant increase of 117% (*i.e.*, 3.5 days to 7.6 days) in the total expected time they would play the game. That is, performance issues may have influenced the game's retention, altering the effect the game could have had in their overall mobility.

3.3.3 Online games effect on mobility

We now study the effects experienced by players of location-based games on human mobility. Analyses using Twitter are associated with the spatial dimension related to places being visited, whereas Carat results focus on app usage, and distances traveled are studied on both sets.

3.3.3.1 Location Based Online Game introduces significant changes to mobility – Carat

To establish this relationship, we split the Carat dataset between different week periods (*i.e.*, weekdays and weekends), and cluster subjects according to their level of engagement with the game. We define engagement based on the number of days a given user is observed playing a certain game, *i.e.*, A [1,21) days, B [21,90) days, C 90 or more days. As discussed previously, this separation between times of the week is aimed at mitigating biases caused by users' routine [126], and the engagement clustering allows us to study the relationship between *dose and effect*. To analyze the changes introduced by the game we contrast Pokémon GO and Clash Royale, a popular game that does not have a mobility component. When comparing to the period before playing Pokémon GO, we note statistically significant (*i.e.*, $p\text{-value} < 0.02$, on weekdays and weekends) boosts in the daily displacements of groups B and C, with over 2 km and 1 km, respectively. Meanwhile, no statistically significant differences were found for Clash Royale players (*i.e.*, $p\text{-value} > 0.09$, on weekdays and weekends). Interestingly, the added changes to B and C *persist after* Pokémon GO use ends.

Table 3.6: Daily movements (in km), per group according to the number of days playing — A: [1,21) days, B: [21,90) days, C: 90 or more days, highlighting statistically significant changes, for Pokémon GO (PG) and Clash Royale (CR). The sample sizes were (995, 1051, 1160) and (257, 317, 230) for (A,B,C) on PG and CR respectively.

Game	Period	A	B	C
CR	Week-day	30.3	30.2	32.5
	Week-end	28.7	26.0	28.3
PG	Week-day	27.3 → 31.2	**28.0 ⇒ 29.9**	**30.6 ⇒ 31.6**
	Week-end	29.3 → 29.4	**28.1 ⇒ 30.4**	**29.6 ⇒ 31.4**

3.3.3.2 Gamers see an increase in daily mobility not observed in non-gamers – Twitter

To reinforce the previous observation and rule out possible sampling biases caused by Carat, we use the Twitter set to compute the total daily displacement Δr from subjects with at least 3 records per day. This constraint aims at ensuring we capture at least part of their daily movements. Furthermore, we split subjects into *gamers* and *control* (*i.e.*, *non-gamers*) in order to demonstrate that any observed change was not equally shared by the entire population. From our analysis, we observe that *gamers* boost their daily movements from 13.1 km to 14.6 (p-*value* = 0.03) on weekdays. Conversely, the *control* group sees a decrease in daily displacements from 16.2 km to 15.9 km during the same period p-*value* = 0.03, which could be possibly explained by seasonality or even significant changes in their usage of Twitter. During weekends, we observed no significant changes for *gamers* while the *control* group saw a significant reduction in their daily movements from 16.7 km to 15.2 km (p-*value* = 0.007).

3.3.3.3 Gamers do not explore entirely new regions – Twitter

Given the various means Pokémon GO requires users to move in order to collect in-game items and achievements, we study whether or not *gamers* travel further away from their typical vicinity. We do so by studying the changes in radius of gyration (r_g) when comparing periods before, during and after the release of the game. Using data from Twitter, we cluster users based on their r_g in intervals of 5 km up to 50 km, with an additional cluster for $r_g > 50$ km. In this analysis, we did not observe any statistically significant change in their distributions of r_g, regardless of the period or group of users considered (*i.e.*, p-*value* > 0.05).

3.3.3.4 Gamers visit new nearby places – Twitter

Knowing that *gamers* maintained their r_g, we next evaluated their changes in the number of total places visited during each studied period. Using data from Twitter, we observed a small but significant boost in the number of unique places when playing the game. That is, while *gamers* saw an increase on average places from 15.4 to 17.4 visited (p-*value* < 0.001), the *control* group saw it go from 18 to 18.9 with no statistical significance (p-*value* = 0.08). Combined with the previous observation, this suggests that *gamers* visited more places in familiar areas.

3.3.3.5 Exploration is stronger for anisotropic gamers – Twitter

To further study the exploration of users during gameplay, we investigate how their isotropy ratio (*i.e.*, spatial regularity of mobility) varies over time. For that, we group *gamers* and *control* on our Twitter data set by their ratios at intervals of 0.2, and analyze them against the Pokémon GO periods. In this study, we note that only highly anisotropic *gamers* experience a significant change in their ratios, that is, *gamers* with polarized visits change into a more homogeneous spread to where they stop. Therefore,

this observation combined with the previous two suggest that *gamers* tend to explore their vicinity while playing the game.

3.3.3.6 Short hops become more prevalent for gamers – Twitter

To understand the changes in mobility brought by Pokémon GO, we compare the distribution of displacements in the periods before and after the game in our Twitter set. For both *gamers* and *control* groups, their distributions follow a truncated Lévy-flight model, however with the observation that the ratio between long and short flights varies differently between groups. Against the baseline in Table 3.5, the powerlaw parameter α changes to greater values for *gamers* ($\alpha_g = 0.35 \pm 0.02$) than for `control` ($\alpha_c = 0.33 \pm 0.02$), *i.e.*, short flights become more prevalent for *gamers* when compared to *control* for the same period.

3.3.3.7 Greater effects on mobility when Pokémon GO improves power consumption

As previously shown, higher engagement drives greater changes to *gamers*' mobility. Therefore, we further analyzed factors which could have altered engagement for Pokémon GO players. From our analysis, we note that significant changes to mobility are only prevalent from Pokémon GO v0.33, which based on their reported changelogs (see § 3.3.2), introduces major changes to improve battery utilization. That is, while initial versions of the game only impacted highly engaged players (*i.e.*, group C, > 90 days), later versions introducing better power management features present significant changes to all groups (*i.e.*, A, B, and C). The implications of these results are twofold: (1) the application architecture may define engagement and, therefore, alter any expected influence on behavior, and (2) divergences in observed effects on behavior from different online games could be explained by the levels of engagement of the studied app [94].

In this section we studied how online games can influence human mobility, beyond its expected periodicity. Mobile games may lead to long-lasting changes which are only observed if the apps contain a mobility aspect to it, like Pokémon Go had. Next, we look into how mobility and contact duration are related.

3.4 Contact and Stay duration as a consequence of mobility

We have so far discussed how mobility and phone utilization are related, including how mobility and network traffic as well as mobility and online gaming are related. These studies allow us to better understand correlations and associations between how mobile users use their smart devices and their movements. We now turn to the study of more fundamental aspects of mobility, namely contact duration between two individuals. The relevance of this topic concerns opportunistic routing [74, 10] as well as epidemics spreading [127, 128, 129, 130, 73]. However, a more accurate model for contact duration remains an open challenge, which we address in this work.

For that, we gather and analyze mobility and sensory data from 71 subjects for 2 months in 2018. Our records include GPS and Bluetooth data of these subjects, allowing us to model the physical encounter between subjects and nearby individuals, including other subjects. Furthermore, through our analysis, we report differences regarding the distributions of stay duration at various places.

3.4.1 Background

We begin by defining a *contact*, *stop*, and *trip* as we use them in this work.

Contact We define a *contact* between any two persons when a Bluetooth signal originating the mobile device of one person is captured by another. We chose this approach because of the small range of Bluetooth emulating well a close contact between persons, particularly for the study of airborne infectious diseases [131, 132].

Stop Given a set of points of interest for a subject, *e.g.*, home, shop or a transit station, we define an extended visit to a Point Of Interest (POI) as a *stop* (or a stay).

Trip Given the descriptions of *stops*, a *trip* is defined as the trajectory between two visits to points of interest. Furthermore, we compute the total length of a *trip* as the sum of all distances between intermediate points of a trajectory, that is $\ell = \sum_t \|\mathbf{x_t} - \mathbf{x_{t-1}}\|$, where $\mathbf{x_t}$ is the location at time t.

3.4.2 Dataset

Starting in April 2018, we collected our data from 71 registered subjects, who gave us *explicit* consent to analyze their traces. For this, we collected data from multiple sensors using the Aware App [133], which manages both the client and server side of the data collection. Our subjects lived in Munich, and are between 20 and 30 years of age. To ensure a reliable and dense set of records, we collected location data (*i.e.*, GPS) and Bluetooth readings at a median rate of one sample every 3 minutes. As our cohort is young, and mostly made of students living in a large and developed European city we note that our observations may not capture all nuances of how an entire population behaves. Therefore, we present some of our observations along with equivalent observations made using the well-known Geolife data set, which include a larger number of subjects.

Spatial Data – GPS The location data were captured by the GPS module of each subject's smartphone, provided as geographical coordinates along with uncertainty values. We note that this reported quality of our data was high, with the 85th-percentile of the uncertainty being 10 meters. This good quality allowed us to accurately estimate stops and trajectories (or trips) of each subject.

Users	Stops	Encounters	Trips
71	19317	12432	18438

Table 3.7: Summary of the data set used.

Contact Data – Bluetooth Similarly, proximity data were captured by the Bluetooth of a subject's phone and further processed to model *contacts*. This was done in a two-step approach, classifying other devices as either *human-held* or *static*. In *step 1*, we worked with the names broadcast by devices by first filtering out non-English/German words, then manually classifying them as mobile or not. These steps ensured that we did not access any personally identifiable information, while also maximizing the coverage of seen devices. Next, similar to the approach taken in § 3.2.2, we further classify devices based on the OUI of their Bluetooth MAC address, further classifying a total of 16902 devices. It is important to note that all other devices which were not classified were discarded to eliminate unwanted uncertainties and biases. After all these curation steps we identified 6500 *human-held* devices which we assume to represent a person.

Even though the assumption that a device represents a person may carry inevitable biases, the analysis of the distribution of contact duration reveals similarities with previous studies, strongly supporting the validity of our set. For the purpose of this study, we analyze contacts at *stops* or *trips*, breaking down those which lasted for both of these states, which accounted for 3.1% of contacts (389 of 12,423). We took this approach to study contacts that happened when the subject was either moving or static, also discarding unwanted biases from personal devices subjects could be carrying themselves. As a consequence, this approach also limited the total duration of possible contacts to a single trip or stop duration.

We report that the distribution of contact duration, either static or mobile, was fitted by a log-normal distribution, with parameters $\mu = 6.67$ and $\sigma = 1.65$ (p-value = 0.002 when compared to a powerlaw). When comparing these results with contacts at *stops*, the major difference corresponds to a larger *shape* parameter (σ), in accordance to previous work where a short-tailed for contacts at stops was observed [134]. A summary of descriptive characteristics of our dataset is presented on Table 3.7.

Supporting set – Geolife

To ensure the validity of our dataset, we support some of our observations with the Geolife data set [135]. This set consists of trajectories from 182 subjects, spanning 4.5 years, and contains GPS points sampled every 5 seconds or every 10 meters. The variety and extensiveness of this dataset allows us to compare the results we found in our set with a bigger group of users as well as an even faster sampling rate.

3.4.3 Stops

We now describe *stops* (or stays) in detail, followed by a model of contacts between individuals at such locations.

Detection of stops

We follow the established method by Zheng *et al.* [136] to detect stops based on GPS trajectories, ensuring robustness and reproducibility. Their method is based on two main parameters: a maximum clustering distance `max_dist`; and a minimum interval of time at a clustered location `min_stop_time` to define a *stop*.

The algorithm iteratively detects stops by: (1) clustering consecutive geographical points until the distance δ between any point is larger than `max_dist`, (2) once a new points no longer fulfills (1), the centroid of an analyzed cluster is marked as a *stop* if the duration of stay at that set of points is longer than `min_stop_time`. For our study and based on the high accuracy of our GPS points, we chose `max_dist` = 10 meters. Furthermore, to choose the best value for the temporal constraint, we graphed the total number of stops detected against values of `min_stop_time` from 1 to 50 minutes, which showed an inflection point between 10 and 15 minutes, leading us to pick the most conservative value for `min_stop_time` = 15 minutes.

Stops data augmentation

To accurately estimate sojourn times from our subjects, we augmented the identified *stops* according to their category. For that, we first identify the "home" location of a subject, then classify all remaining *stops* with a combination of location APIs.

Identifying home The accurate identification of a subject's "home" is paramount to the study of human mobility for the central role this location has on people's mobility [11, 137]. This identification is done by assigning "home" to the *stop* accounting for the highest number of visits between 7pm and 7am [137]. We then proceed to discard these locations as well as any Bluetooth device ever observed there from any further analysis. This exclusion is done to focus our analysis outside people's homes, where they are not likely to have any control over whom they might encounter.

Identifying the remaining stops For all other *stops*, we used 4 location APIs: Google Places[4], Tomtom Places[5], Foursquare Places[6], Here Geocoding and Search[7]. For a given geographical coordinate, these services provide a set of nearby POIs, from which we pick the nearest suggestion and discard any option that is more than 10 meters away. This heterogeneity of services was used to ensure the highest possible coverage of searched points, with which we identified 57% of *stops*.

[4]https://developers.google.com/places
[5]https://developer.tomtom.com/products/places-api
[6]https://developer.foursquare.com/docs/places-api/
[7]https://developer.here.com/documentation/geocoding-search-api

Categories for places To better model the sojourn times, we later use the categories for places from the previous step. The identified categories of POI were: apartment/residence, bank, bar, company/office, entertainment (*e.g.*, museum, art gallery), gas station, gym/sports facility, health facility (*e.g.*, hospital, clinic), hotel, library, religious center, restaurant, salon, shop, supermarket, theater (including cinemas), transport station (*e.g.*, train, bus), and university.

Analysis of stops duration

We now present the results for *stop* duration, or *sojourn time*.

Uncategorized stops follow a powerlaw Without any further classification, the distribution of stop duration is well described by a powerlaw with $\alpha = 2.13$ (p-value < 0.001 to a log-normal). For comparison, the same analysis on the Geolife set yields comparable results, with $\alpha = 1.98$ (p-value < 0.001 to a log-normal), as depicted in Figure 3.9. This long-tailed distribution is often associated to preferential attachment, in which a subject prefers returning to previously visited locations. As a consequence, most places are rarely visited while few places are often observed.

Categorized stops reveal visit patterns When considering the categories of places in the analysis of stop duration, some categories present a long-tailed powerlaw distribution while others present an exponentially-tailed log-normal distribution. In the latter case, this observation suggests that the underlying process is bounded by the available resources, such as money, time or space available to be used. Furthermore, the powerlaw distribution for stop duration could be the result of a mixture between stops of different types, as Kai *et al.* [44] previously demonstrated for a mixture of log-normal distributions forming a powerlaw.

Type of stop according to stay duration A shared characteristic of *stops* best fitted by a log-normal distribution is the typical "lack of time" constraint to enter or leave these places, such as bars, restaurants and gyms. Therefore, we refer to these as *time-unbounded-stops*. In contrast, places where a visit typically follows a schedule, such as offices, hotels and transport stations, where best characterized by a powerlaw, which we refer to as *time-bounded-stops*.

3.4.3.1 Contacts Characterization at Stops

Contacts are log-normal, even when controlling for distance from home At *stops*, the distribution of contact duration is best fitted by a log-normal distribution, as depicted in Figure 3.10. Additionally, we note that this distribution remains the same even when we vary the distance from a user's home at up to 1 km, between 1 km and 100 km, and even above 100 km.

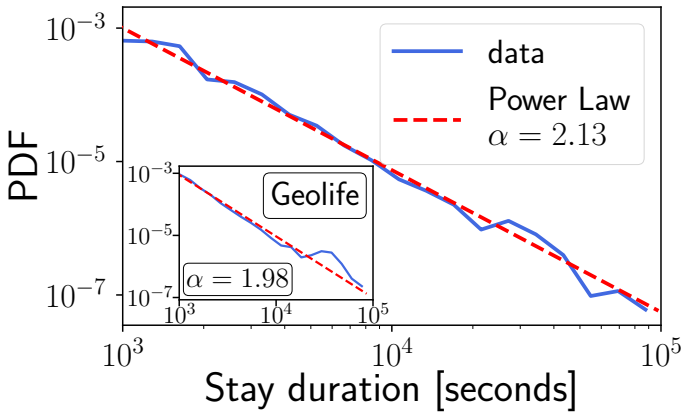


Figure 3.9: Overall stop duration follows a **power-law** distribution.

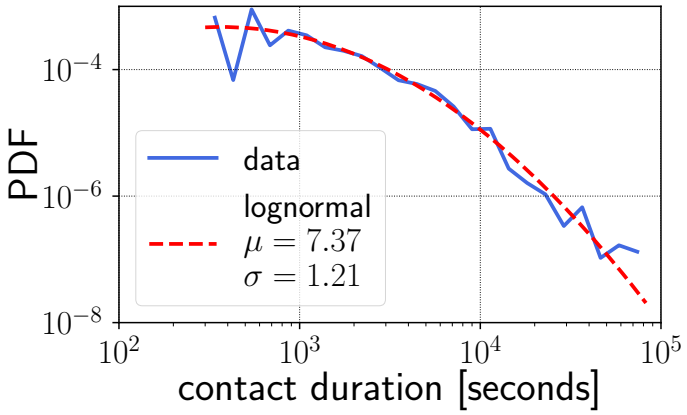


Figure 3.10: Contact duration at stops follows a **log-normal** distribution.

Contacts at different types of stop Using the classification proposed above for *stops*, we observe similar distributions for contact and stop duration. That is, at *time-bounded-stops* we observe a powerlaw distribution for contacts ($\alpha = 2.21$, p-value $= 0.03$ to a log-normal), whereas we see a log-normal at *time-unbounded-stops* ($\mu = 7.6, \sigma = 0.99$, p-value $= 0.04$ to a power-law). This could be explained by individuals spending longer periods of time at *time-bounded-stops*, allowing for longer contact periods.

3.4.3.2 Model of contacts during stops

As the availability of inter-personal contact data is not commonplace, we devise a simple model for contact duration from stay duration. This could be relevant in settings where data protection policies prevent the sharing of individual data on contact times, but aggregate stay duration is available. Our proposed model could be used to study the

dissemination of information opportunistically, and to predict the spread of infectious diseases.

As we have previously discussed in our results, firstly, we define the visit duration probability y as a powerlaw, *i.e.* $Pr(y) = Cy^{-\alpha}$, where $C = (\alpha - 1)x_{\min}^{(\alpha-1)}$, and α is the defining coefficient of the distribution. Then, as the contact duration x is described by a log-normal distribution, we can write $e^x \propto N(\mu, \sigma)$, where $N(\mu, \sigma)$ is a normal distribution defined by μ and σ. To avoid the non-trivial estimation of $N(\mu, \sigma)$ we can approximate its probability density function with a uniform distribution.

This non-parametric approximation yields a constant loss-function in the interval of a stay duration (*i.e.*, from 0 to y) w.r.t. its original distribution P. This observation emerges from the KL-Divergence between any target distribution P being approximated by a Uniform distribution U, in the interval (a, b) (or simply n, where $n = b - a$) as in Equation 3.1. In this case, the final distance between the two distribution functions is defined only by n and $H(P)$ (*i.e.*, the entropy of the target function).

$$\begin{aligned} D(P||U) &= \sum_{i}^{n} P(X_i) \log_2 \left(\frac{P(X_i)}{U} \right) \\ &= \sum_{i}^{n} p_i \log_2 \left(\frac{p_i}{1/n} \right) \\ &= \log_2(n) + \sum_{i} p_i \log_2(p_i) \\ &= \log_2(n) - H(P) \end{aligned} \tag{3.1}$$

With these observations, we can re-write x as $e^x \propto 1/y$, and therefore relate x and y as $Pr(x)\, dx = Pr(y)\, dy$, or equivalently $dx \propto e^x\, dy$. Substituting, we get $Pr(x) \propto C\, e^{\alpha-1}$. Additionally, a random variable Z is described by a log-normal if it has the form $Z \sim e^{\mu+\sigma x}$ and if x is normally distributed. By comparing this log-normal definition with $Pr(x)$ inferred above, we can write $\mu \approx \ln(\alpha - 1)x_{\min}^{\alpha-1}$ and $\sigma \approx \alpha - 1$. Finally, from our analysis previously discussed, we take $\alpha = 2.13$ (from stay duration) and estimate $\hat{\mu} = 7.80$ and $\hat{\sigma} = 1.13$, which are close to the actual values $\mu = 7.37$ and $\sigma = 1.21$ (from the observed contact duration).

Using our proposed model, we vary the scale parameter α for stay duration and graph the resulting distributions of contact duration in Figure 3.11. Interestingly, our model highlights how shorter visits yield a smaller probability of shorter contacts while increasing the probability of longer ones. That is, a larger α (*i.e.*, overall shorter stays) leads to larger $\hat{\mu}$ (*i.e.*, the distribution shifts to the right) and $\hat{\sigma}$ (*i.e.*, the spread of the distribution gets larger). It is important to note that, this change in contact duration does not relate directly to the frequency of contacts, but it rather relates the prevalence of stays of different duration. That is, if people were to stay shorter periods of time at places, this naturally increases the probability of longer contact duration.

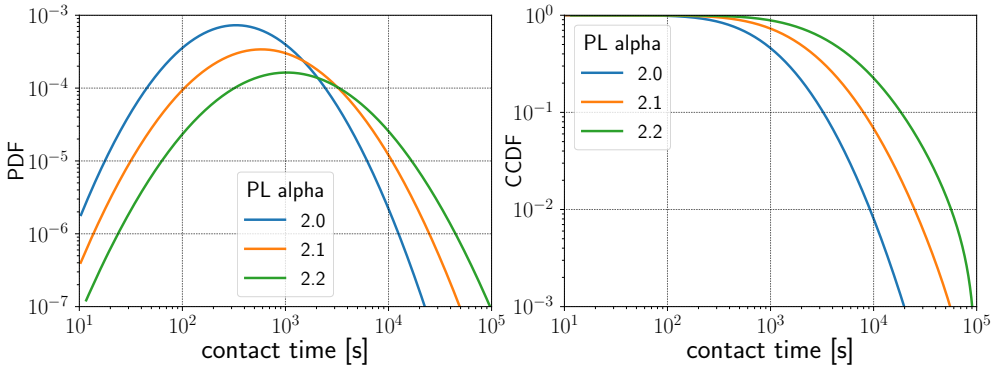


Figure 3.11: Distribution of modeled contact duration for different values of the stay duration parameter (*alpha*). Larger values of *alpha* for stay duration indicate *higher* probability for shorter stays, leading to an increase in the probability of long-term contacts as short-term meets become less often.

3.4.4 Trips

We now present our analysis of *trips*, including our observations that contact duration, while on the move, is best described by a Weibull distribution.

Detection of trips

To ensure robustness in detecting *trips*, we use a three-step approach. First, we discard any trajectory that does not begin *and* end in a previously identified *stop*. Second, we apply a time constraint to filter out trajectories which contain any two points recorded more than 1 hour apart, eliminating trips containing large gaps. Lastly, we apply a spatial constraint to filter out trajectories in which any gap accounts for more than 50% of the trip total length, ensuring the continuity of traces during *trips*. On our dataset, this approach yields 2512 *trips* which we analyze next.

Trip Duration and Total Length

From our analysis, *trip* duration was best fitted by a log-normal distribution p-value = 0.02 to a powerlaw), depicted in the left panel of Figure 3.12. On the inset of this same panel, we also present a similar observation done with the Geolife set (p-value < 0.001 to a powerlaw). The exponential tail of these distributions, instead of a long tail, could be explained by a decrease in the expected population density in suburban areas along long trajectories [68]. These observations help ensure the robustness of our data and approaches to detect *stops* and *trips*.

Similar to results by Alessandretti *et al.* [138] (N=850, using GPS points) and our own analysis of the Geolife set, trip length is best fitted by a log-normal distribution in our data, graphed in Figure 3.12. For this same analysis, a powerlaw fit shown in dotted grey yields $\alpha = 1.22$, however with a statistically significant lower log-likelihood (*i.e.*,

p-value = 0.009). These results are in stark contrast to previous work using call detail records [11, 36], which showed powerlaw as the best fit.

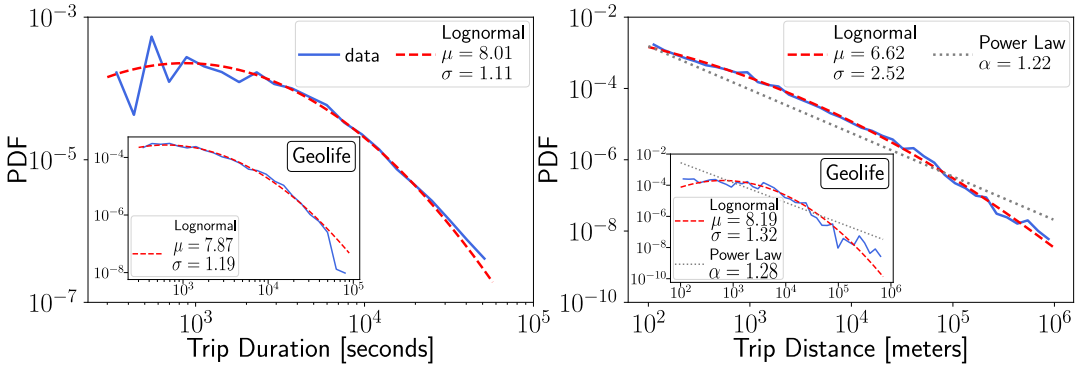


Figure 3.12: Both trip time duration and length are best modeled by a log-normal.

Model of contacts during trips

When considering all *trips*, contact duration was not significantly better fit by any of the considered distributions (*i.e.*, p-value > 0.05 for any pair of distributions). A better fit emerged when controlling for the trip length. This approach unveiled an interesting link between trip length and contact duration, as shown in Figure 3.13.

We perform this analysis by observing the changes in contact duration with different parameter values for trip length. With longer trip lengths, λ (*i.e.*, the scale of the distribution, proportional to its expected value) displays a bi-modal shape. We conjecture that this segmented shape captures the tendency people have to take crowded public transport (bus, train, airplanes) when covering longer distances. As an alternative, people could either walk or drive their own cars, yielding less (and shorter) contacts with other people.

Similarly, β (*i.e.*, the shape or dispersion of the distribution) decreases from $\sim$1.2 to $\sim$0.5 as the trajectory length gets shorter. With that, it is important to note the interpretation of different values of β: when $\beta > 1$ the probability of contact duration drops exponentially, resulting in extremely low probabilities for higher durations, and when $\beta < 1$ the function behaves like a long-tailed distribution, resulting in a few large contact durations. We also conjecture this result to be due to a choice for a preferred means of transport. That is, when traveling longer distances on public transport people are likely to be found next to some long distance travelers.

Unlike with *stop* duration, modeling changes in *trip* duration requires a deeper understanding of the reasons people move larger distances [137]. For example, daily *trips* from a subject's home to their office tends to be fixed, or to a bar or gym following someone's intentions. Therefore, to thoroughly capture variations in *trip* length distribution and their resulting contact duration by, *e.g.*, lockdown measures, a recent dataset with equivalent spatial and temporal density is required. Therefore, this possible effect of lockdowns was not consider in our studies.

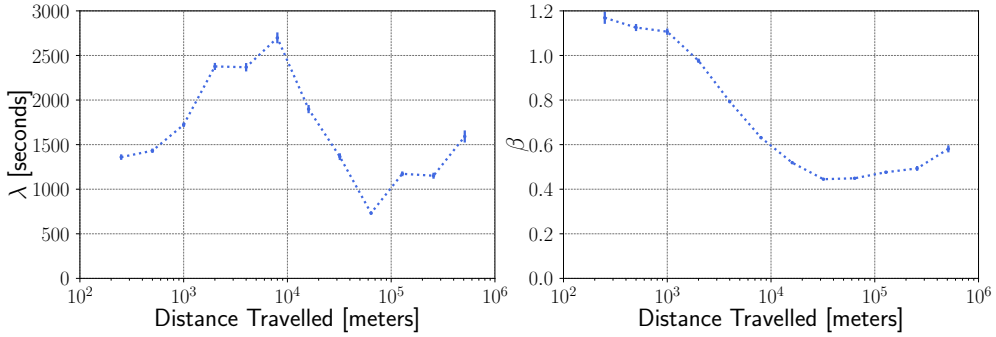


Figure 3.13: Variation of the Weibull parameters as a function of distance traveled.

3.5 Mobility Networks and Epidemic Forecasting

The ample disruption caused by the COVID-19 pandemic highlighted the relevance of better understanding the mechanics of epidemics spread in large populations [139]. This understanding may be applied to devise better control strategies such as immunization through vaccination and containment strategies for future outbreaks [140, 141, 142, 143]. However, a seemingly simple process – one individual infects another – becomes a complex phenomenon when considering human society and the intricacies of contacts as discussed in the previous section. These interactions yield characteristics typically observed in human networks such as the formation of cliques and scale-free networks, with many nodes having few connections and few nodes being extremely well-connected, all having a significant impact in how diseases may spread through these networks [144, 145, 146].

Research in mobility and network epidemics allows us to study these complex problems. For that, we can simulate outcomes of an epidemic outbreak using temporal networks resulting from people's contacts. Furthermore, learning models can be applied to predict states of nodes, for example, in the absence of data about all nodes affected by an outbreak. However, most commonly used prediction methods use feature vectors as input and not a temporal graph inferred from contacts.

To tackle these limitations, current learning methods use node embeddings to translate a temporal network of contacts into feature vectors. This is done by representing network nodes as low-dimensional vectors without losing information about the structure of the graph. That is, structurally similar nodes will result in nearby feature vectors, preserving node similarity. For disease spreading predictions, the quality of node embeddings is defined by their ability to capture information about *when* and *who* is infected.

Two relevant concepts for such predictions are *homophily* and *structural similarity*. The former describes nodes that are connected for sharing some kind of similarity, and is often the focus of disease spreading models as it captures direct interactions between individuals. Additionally, *structural similarity* describes nodes with similar positions in the network structure and, along with *homophily*, will have different effects on a network [147].

For our analysis, we use `node2vec` as our node embedding method, which encodes a graph's topology using random walks [147], *i.e.*, this approach converts an adjacency matrix into matrices of numbers by randomly sampling nodes and their neighbors. To control these random walks, the algorithm can either bias *inward* exploration, *i.e.*, preserving (or favoring) homophily, or bias *outward* exploration, *i.e.*, preserving structural equivalence.

Current methods, however, tend to favor homophily in their analysis of epidemics due to its properties of capturing person to person interactions. We address shortcomings of these approaches by comparing predictions of node states when favoring random walks with an inward, neutral, and outward orientation. Our study reveals that best results are achieved with *outward* oriented explorations, suggesting a greater importance of structural equivalence during an infectious outbreak. That is, when trying to predict who is more likely to be infected next during an epidemic outbreak, distant neighbors have a more significant predictive power than previously observed. These distant neighbors, when sharing similar near network structure (*e.g.*, two university professors in distinct departments), should be also considered during a prediction exercise. Our results highlight that structurally equivalent nodes could be at similar risks of infection, in spite of possibly not being directly connected. We believe these observations could support future contact tracing efforts during an outbreak of an infectious disease.

3.5.1 Methods

3.5.1.1 Structural Equivalence and Homophily

In sociology, the term *homophily* is defined and studied as cohesion or strong ties in society, whereas the term *structural equivalence* is often linked to weak ties. Research has shown that, in certain circumstances, weak ties may be more important than strong ties [148], highlighting the relevance of studying network structures originating from human contacts. In network science, when studying the similarity between nodes, *homophily* is represented by the common neighbors of two nodes, and structural equivalence is characterized by the affinity in the structural position any pair of nodes have, even if they are not directly connected. In the study of certain processes, including epidemic dynamics, a mixture of these two concepts needs to be preserved and balanced by the embedding algorithm for an accurate representation of society. For that, we chose `node2vec` as our embedding algorithm.

Node embedding is a technique for representing a graph in a predictive model. By transforming nodes into dense, low-dimensional vectors, node embedding enables the use of traditional machine learning algorithms that are designed for vector inputs. This allows for more accurate predictions and better performance on downstream tasks such as node classification and link prediction. Additionally, node embedding can capture the structural properties of the graph, such as community structure and centrality, which can further improve the predictive power of the model. In summary, node embedding is a critical step in graph analysis and modeling that can significantly enhance the accuracy and efficiency of predictive models.

3.5.1.2 `node2vec`

The `node2vec` algorithm balances *homophily* and *structural equivalence* through a series of random walks [147]. In this method, the number of walks and their respective lengths are defined as hyperparameters. More importantly, `node2vec` has two additional hyperparameters controlling the bias in these exploratory random walks, that is, these parameters control whether the exploration should remain close (*i.e.*, inward) or far away (*i.e.*, outward) from the starting node. Intuitively, inward explorations are similar to the search tree of a breadth first search, whereas outward explorations are similar to a depth first search.

In this algorithm, the `p` parameter defines the probability that the random walk will return to a previously visited node, yielding higher similarities between nodes in the resulting node embedding. Additionally, the `q` parameter, also referred to as in-out parameter, forces the random walk to reach nodes that are further from the starting node. The authors of the `node2vec` method [147] suggest that mode inward random walks favor *homophily*, whereas more outward probing conserves *structural equivalence*.

In our work, temporal networks, from real-wold mobility traces or synthetic, were embedded with `node2vec` using different hyperparameters (see Table 3.8). We observe that the length and number of random walks taken does not significantly influence accuracy predicting how an epidemic evolves. Therefore, for our remaining analyses, we set the number of walks to 10 and their length to 80. That is, to generate the embeddings from each temporal network, we sample the structure of the nodes with 10 random walks, that will traverse 80 randomly selected nodes.

We compared prediction accuracies for 5 values of inward and 5 values of outward exploration, resulting in 100 different pairings of `p` and `q`. The resulting embeddings were then used to predict node labels from 250 SIR-simulations per network (see Section 3.5.1.5), for which we used logistic regression with L2 regularization for its simplicity and stability.

Moreover, in our work, `node2vec` may not be able to capture the full extent and purpose of social interactions. In this work, we assume that two nodes being nearby each other constitutes a contact in which the infectious disease could spread. This may limit the predictive power and generalizability of the embedding to new scenarios or populations.

3.5.1.3 Datasets

For this study, we used a total of 24 datasets, with six coming from real world measurements and all are publicly available and were collected in other research studies with explicit consent from all participants or their responsible guardians. One of these sets (the Reality Mining) was collected by the MIT Media Labs and contains location information along with Bluetooth readings from nearby devices [149]. The other five real world sets were collected by the SoctioPatterns project and include proximity readings from RFID readers [150]. In all our analyses, all methods were performed in accordance

Name	Place	Year	Participants	Duration	Sampling Method
InVS15	work office	2015	232	2 weeks	RFID
LH10	hospital	2010	81	3 days	RFID
LyonSchool	primary school	2009	242	2 days	RFID
realitymining	university	2004	100	9 months	Bluetooth
SFHH	conference	2009	403	2 days	RFID
Thiers13	high school	2013	326	1 week	RFID

Table 3.8: Metadata of the data sets used

with the relevant ethical and legal guidelines and regulations. Detailed information about the datasets can be found in Table 3.8.

From these datasets containing timestamped contacts, we derived corresponding temporal networks where participants are nodes and each registered contact is an edge. We binned data in intervals of 10 minutes, which achieves a good balance between a minimum time for an infectious disease to spread and having enough time steps to conduct our study.

Additionally, we used 18 random temporal networks with node degrees sampled from a binomial distribution in one half and powerlaw in the other half. These distributions were chosen for their simplicity, in the case of binomial, and for their similarity with other real world networks, with powerlaws [151].

3.5.1.4 Network Representation

To derive the temporal networks, time steps of each network were connected into a static supra-adjacency network. Nodes are identified by pairs (k, t), where k is a node and t the time step. If k is infected at t, it is still likely to be infected at $t + 1$. To use this temporal dependency in the prediction, the time steps of the network are interconnected, *i.e.*, node (i, t) is connected to (i, t+1). Furthermore, if there is a contact between i and j at time step t, there exists an edge from (i, t) to (j, t+1) and from (j, t) to (j, t+1). Additionally, only active nodes are considered, reducing the total number of nodes that need to be embedded. These are nodes (i, t) where i had at least one contact at time t. Inactive nodes are then deleted and their incoming edges are rerouted to their next active future self, as previously done by Sato *et al.* [109]. Finally, we converted all 24 datasets into supra-adjacency networks, where Table 3.9 describes the sizes and densities of the real world networks, with artificial networks described on Table 3.10.

3.5.1.5 SIR simulation

As real world data on infection outbreaks are hard to obtain and would not be available for privacy concerns, we used a compartmental model to simulate how the disease would spread in our studied networks. Compartmental models are commonly used mathematical tools to define and run such simulations in order to generate reference values to be predicted, *i.e.*, states of nodes such as susceptible or infected. In our study, we model the spread of an infectious disease through our temporal networks using an SIR model,

Table 3.9: Metadata of the dynamic networks: $|V_A|$ active nodes, $|V|$ participants, $|T_A|$ active timesteps, $|T|$ timesteps and $|E|$ number of edges

Name	$\|V_A\|$	$\|V\|$	$\|T_A\|$	$\|T\|$	$\|E\|$
InVS15	22133	217	698	1656	18386
LH10	5219	76	359	477	7580
LyonSchool	18641	242	116	217	45008
realitymining	21613	102	980	1025	50897
SFHH	10693	403	128	191	17003
Thiers13	32015	327	246	606	34920

Table 3.10: Artificial Networks: node degree distribution, $|V_A|$ active nodes, $|V|$ participants, $|T_A|$ active timesteps, $|T|$ timesteps and $|E|$ number of edges

Distribution	$\|V_A\|$	$\|V\|$	$\|T_A\|$	$\|T\|$	$\|E\|$
binomial	44925	1000	100	100	29912
binomial	45105	1000	100	100	29951
binomial	45112	1000	100	100	29998
binomial	17926	100	400	400	12080
binomial	17395	100	399	400	11644
binomial	17443	100	400	400	11718
binomial	2116	10	275	300	2281
binomial	2233	10	290	300	2476
binomial	2050	10	286	300	2154
powerlaw	45057	1000	100	100	29913
powerlaw	45149	1000	100	100	30003
powerlaw	45027	1000	100	100	29956
powerlaw	17879	100	400	400	12076
powerlaw	17501	100	399	400	11646
powerlaw	17390	100	400	400	11709
powerlaw	2110	10	275	300	2256
powerlaw	2209	10	290	300	2493
powerlaw	2060	10	286	300	2142

which starts with a defined number of *infected* (I) nodes. Next, at each following time step, infected nodes may infect other *susceptible* (S) connected nodes with probability α (infection rate), and it may also become a *recovered* node (R) with probability μ (recovery rate) and will never change state again.

Note that, in this model, at each step a node will be assigned only one single state: S, I, or R. That is, if nodes are "infected" they are contagious and may spread the disease to other, "susceptible" means they are still vulnerable, and "recovered" means they are either healthy and immune or are dead no longer playing a role in the disease spread mechanism. Note however, that while in the SIR model states can only go from S to I then R, prediction models do not have this limitation as they are built agnostic of the

Table 3.11: Parameters for the SIR simulations

(infection rate, recovery rate)
(0.13, 0.002)
(0.25, 0.002)
(0.13, 0.005)
(0.25, 0.005)
(0.25, 0.007)

semantics or ordering of these labels. For each network, we ran SIR simulations with 5 different parameters set (α, μ), which can be found in Table 3.11. These sets were chosen around ranges in which there was an observed outbreak lasting for the entirety or majority of the simulation.

3.5.1.6 Evaluation Metrics

For our evaluation metric, we chose the f1-score for its robustness in evaluating prediction accuracy, by balancing true and false positives or negatives. Since the size of classes can be strongly imbalanced (*e.g.*, many more S than I nodes at a given time), we include two alternatives to further expand the f1-score to multiple labels, namely the *micro f1-score*, which weighs all classes the same, and the *macro f1-score*, that evaluates scores separately for each class then reports the average over all classes. To evaluate aspects of the spreading process of an epidemic outbreak, we considered different disease-specific metrics:

- **End outbreak size**: total number of infected or recovered people, which captures the extension of the outbreak.
- **Mean prevalence**: the expected number of infected people at any time step.
- **Peak prevalence**: the maximum number of infected people at one time step, which can help estimate hospital capacity to care for infected people in "the worst case scenario".
- **Peak prevalence time**: the time step of peak prevalence, which is commonly associated with how aggressively the disease spreads.
- **Mean incidence**: the expected number of nodes changing their state at any time step, which is indicative of the disease spread rate.

All metrics are evaluated as the difference between simulation and prediction, and to establish a meaningful comparison across networks results are reported as percentage, either of participants or number of time steps.

q \ p	0.01	0.3	0.5	1.0	10.0	50.0	80.0
0.01	68.60%	70.65%	70.97%	70.80%	70.82%	71.01%	70.93%
0.3	63.65%	67.67%	68.75%	70.29%	69.02%	70.40%	70.33%
0.5	60.93%	69.27%	70.05%	70.08%	70.51%	70.43%	70.73%
1.0	60.46%	67.66%	68.95%	69.74%	70.15%	69.41%	70.44%
10.0	55.19%	57.96%	62.22%	64.09%	65.51%	67.09%	67.44%
50.0	58.03%	55.86%	57.56%	58.81%	60.32%	61.53%	64.33%
80.0	58.13%	56.36%	57.26%	60.29%	59.12%	60.93%	62.43%

Table 3.12: Average micro f1-score for different values of `p` and `q`, with outward exploration marked in pink and inward exploration in white

q \ p	0.01	0.3	0.5	1.0	10.0	50.0	80.0
0.01	-1.15%	0.91%	1.22%	1.06%	1.08%	1.26%	1.18%
0.3	-6.09%	-2.07%	-0.99%	0.55%	-0.73%	0.66%	0.58%
0.5	-8.81%	-0.47%	0.31%	0.34%	0.77%	0.69%	0.99%
1.0	-9.28%	-2.08%	-0.79%	0.00%	0.40%	-0.34%	0.69%
10.0	-14.55%	-11.78%	-7.53%	-5.65%	-4.23%	-2.65%	-2.30%
50.0	-11.72%	-13.88%	-12.18%	-10.94%	-9.42%	-8.22%	-5.41%
80.0	-11.62%	-13.38%	-12.48%	-9.45%	-10.62%	-8.82%	-7.31%

Table 3.13: Difference in micro f1-score to the unbiased embedding

3.5.2 Results

Embeddings with outward exploration outperform those with inward exploration in predicting epidemic dynamics, *i.e.*, structural similarity has a stronger impact on predictions than a node's exact neighbors, or a node's role in a network could be more relevant than who their peers are for determining disease outbreak outcomes.

We compare predictions using either inward exploration, where nodes with common neighbors are considered similar, or outward exploration, where nodes surrounded by an equivalent structure of neighbors are considered similar. Next, we evaluate our results using the f1-score, which balances true and false positive and negative predictions in all time steps of the outbreak (see Methods § 3.5.1), which uses 50 simulations for each combination of dataset and SIR parameters, out of which we report the average results with their corresponding confidence intervals.

In our results, we note that by prioritizing *outward* explorations (*i.e.*, higher return parameter `p`, or a lower in-out parameter `q`), we achieve a better f1-score, depicted in Figures 3.14(a) and (b). Against neutral values, outward exploration achieved an average improvement in the micro f1-score of 0.01393 (95% CI: -0.033 to 0.061), and

Table 3.14: Mean and maximum scores for inward/outward parameters

	micro f1-score	macro f1-score
neutral	69.4%	52.9%
mean outward	70.8%	54.8%
mean inward	60.4%	40.2%
max outward	71.0%	55.1%
max inward	67.5%	50.0%

inward exploration reduced the f1-score by 0.090211 (95% CI: 0.037 to 0.143). Further results are presented in Tables 3.14, **??** and **??**. The best scores come from the set of parameters with high `p` and low `q`. This focuses the network structure representation on structurally equivalent nodes rather than their immediate neighbors, resulting in improved prediction quality.

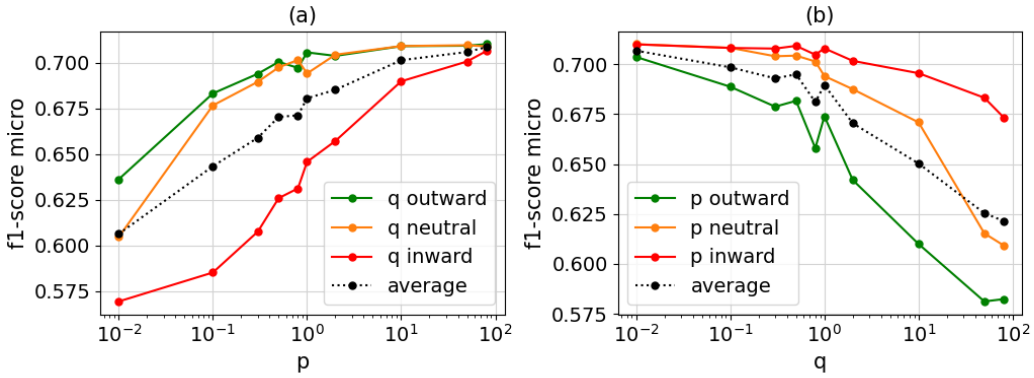


Figure 3.14: Comparison of inward, outward and neutral parameters

Additionally, in certain time steps some labels (S, I, or R) appear less frequently in our simulations and are therefore harder to predict, resulting in a lower macro f1-score than the micro f1-score (see Figure 3.15). Nevertheless, the prediction accuracy score increases similarly to the micro f1-score by 0.0185 (95% CI: -0.029 to 0.066) for outward and decreases by 0.1276 (95% CI: 0.091 to 0.164) for inward exploration (see Table 3.14).

Our results reveal that greater outward exploration leads to better prediction of disease-specific metrics, with the most significant improvements seen in predicting final outbreak size and peak prevalence time (see Figure 2). Furthermore, mean prevalence prediction has the least improvement, with a difference of 0.0284 percentage points (95% CI: 0.027 to 0.029). The effect of outward exploration can also be seen in the incidence, with an improvement of 0.0134 percentage points (95% CI: 0.011 to 0.015) (see Figure 3.16). In our simulated parameters, the improvement of prediction accuracy seems to be the strongest early in the epidemic, showing a higher macro f1-score for the first 10% of time steps (see Figure 3.17). This effect is however not reflected in the micro

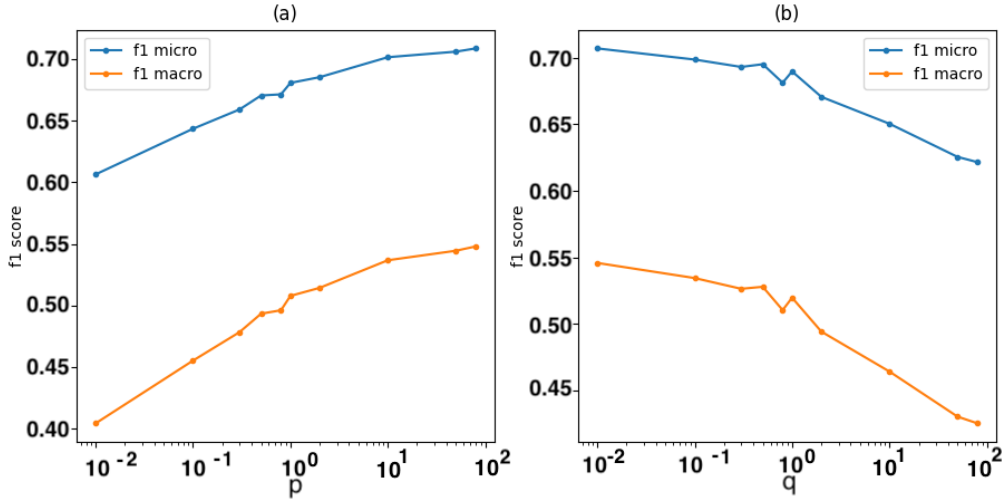


Figure 3.15: Comparisons of micro and macro f1-score for different values of `p` and `q`

f1-score, where the difference between inward and outward exploration remains the same in different phases of the epidemic (see Figure 3.18).

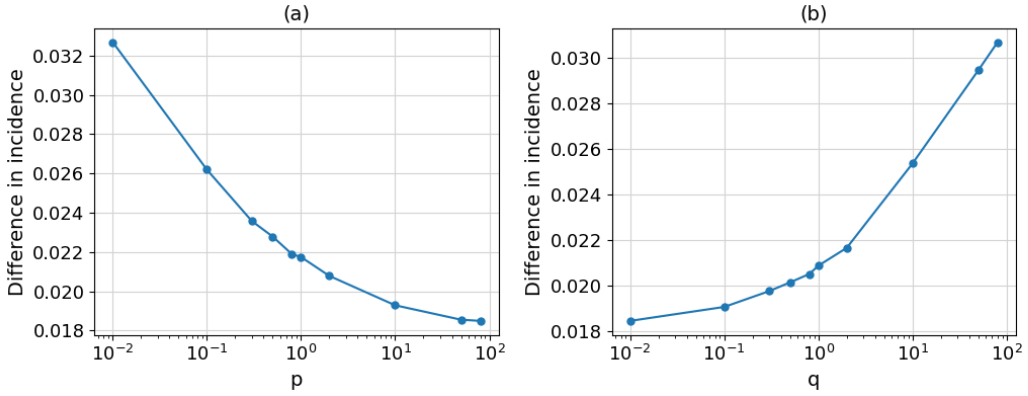


Figure 3.16: Comparisons of the mean incidence for different values of `p` and `q`

In the modeling and prediction of an infectious disease spread, two nodes are considered *similar* if their label (*i.e.*, predicted outcome) is likely to be the same, meaning they will get infected and recover at the same time. Additionally, a node's state will depend on its neighbors as it is likely to be infected at the same time as its neighbors as they infect each other. Therefore, knowledge about the state of any neighboring node could still help prediction accuracy, supporting the relevance of considering homophily when studying disease spreading.

Representing nodes in a temporal network over time steps may not be optimal for predicting disease spread. Typically, node i at time t is connected to node i at $t+1$ in

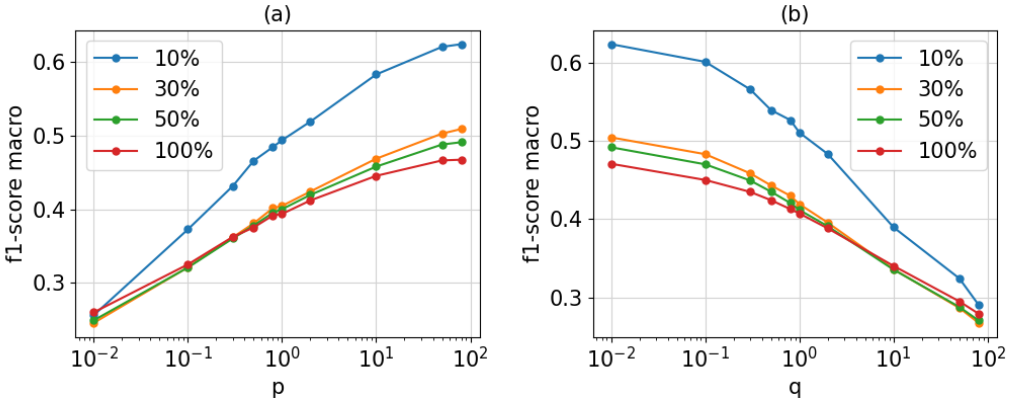


Figure 3.17: Comparisons of macro f1-scores for the first 10%, 30% and 50% of time steps for a random sample of 150000 predictions

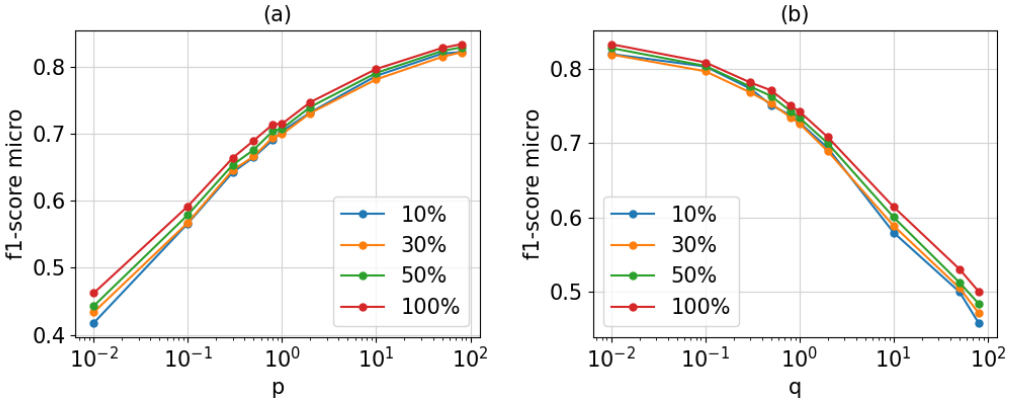


Figure 3.18: Comparisons of micro f1-scores for the first 10%, 30% and 50% of time steps for a random sample of 150000 predictions

a supra-adjacency network, mimicking reality as these nodes represent the same person at different times and are likely to have the same state.

But in contrast to predicting interests, predicting the dynamics of an epidemic outbreak does not necessarily rely on node similarity causing the connection. Instead, two nodes become similar *due* to their contact. In the corresponding temporal network, this creates a time delay where the two nodes are only similar for one time step after contact.

Our results show that structural equivalence can indicate a person's infection time. People in the same structural position are likely to get infected around the same time. Additionally, communities can also play a crucial role in disease spread. If one community gets infected and has two neighboring communities, nodes that act as bridges between these communities are likely to be infected first, followed by central nodes in the neighboring clusters, and then nodes at the edges. Knowing the state of a node in a neighboring cluster can allow us to forecast the state for structurally equivalent nodes in the other clusters.

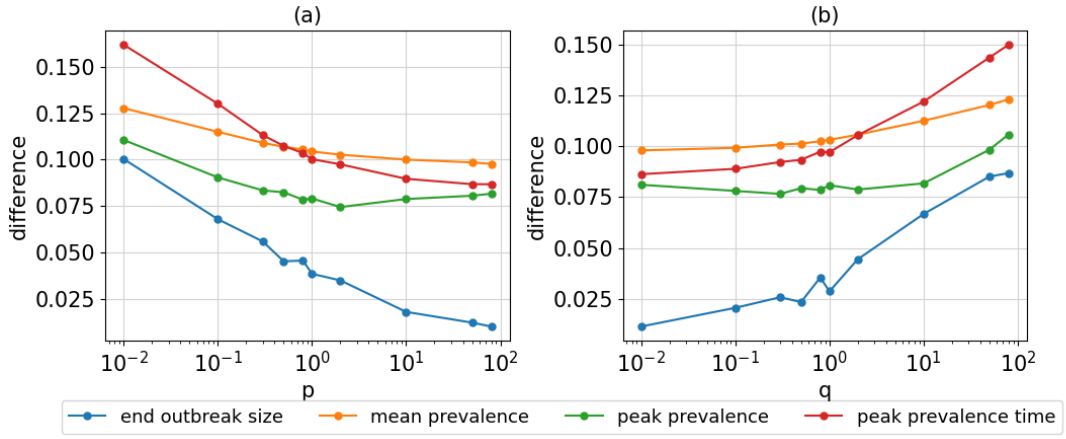


Figure 3.19: Difference in epidemiological metrics between prediction and simulation

Balancing structural equivalence and homophily, or outward and inward exploration in a random walk, helps in understanding the disease spreading process. Oftentimes, predictions follow the homophily approach, tracking the disease from one person to another. However, our results suggest that predictions are most accurate with outward-oriented parameters and that considering the information about structurally equivalent nodes in the training set enhances a node's state prediction.

3.5.3 Relevance of the results

Node embeddings are critical for modeling graph behavior, including infectious disease spread. For that, `node2vec` effectively captures structural equivalence, allowing us to study the impact of homophily and structural equivalence on human contact networks. The spread of diseases like COVID-19 is driven by close contact between individuals [152, 153], with frequency and timing of interactions determining outbreak outcome [154, 155]. Our results suggest a person's network position also indicates infection timing, and the encoding of this information in a node embedding can be controlled via the inward and outward exploration parameters of the chosen node embedding algorithm (*i.e.*, `node2vec`).

Improved prediction accuracy can aid in preventing disease spread by informing countermeasure design. Structural equivalence in contact networks reflects people's societal roles. Our proposed approach may also reveal early infected groups with limited temporal contact information. For example, a high rate of infections among hospital staff or schoolchildren in one community suggests structurally equivalent groups in similar communities may also become infected.

Our results emphasize the need for temporal contact data in studying disease spread. To prepare for future pandemics, privacy-protected contact tracing efforts should be established. Human mobility is highly predictable [26], making sparse contact sampling prior to outbreaks representative. Analyzing structurally similar nodes can indicate an

increase in infections. Additionally, further validation is needed through analysis of actual outbreak scenarios, not just simulations.

3.6 Discussion

This chapter reveals the close relationship between mobility and various aspects of human behavior. We correlated individual movements with network traffic and mobile phone usage, enabling us to construct empirical models that explain and predict mobility-related phenomena. Additionally, we investigated how daily mobility influences interpersonal contacts, leading to distinct patterns, and how a contacts network could facilitate more accurate infectious disease predictions.

Thus, we tackled the initial research question (**RQ1**) regarding studying and establishing a correlation between human mobility and phone usage. Our methodologies and findings indicate that this association stems from mobile devices' ability to detect mobility and their capacity to alter predictable patterns through their services.

The location and proximity data obtained from mobile devices provide us with temporal and spatial granularity, enabling us to track users across time and space. Analysis of these data reveals patterns, which contribute to the development of robust predictive models. However, the pervasiveness of wireless networks and mobility-related online services can disrupt the previously established regularity in human mobility [26].

Incorporating additional features into the modeling of infectious disease spread can further improve the accuracy of predictions. For example, demographic and behavioral data can be used to identify individuals who are more likely to become infected or transmit the disease to others. In addition, environmental factors, such as air pollution or temperature, can also impact the spread of certain diseases. By integrating these features into the node embedding algorithm, we can capture the complex interplay between different factors that contribute to disease spread.

Despite the benefits of this data and online services, privacy concerns often arise [156]. Consequently, new regulations like GDPR and CCPA have been implemented. To comply with these rules while studying human mobility, privacy-preserving methods are necessary. In the subsequent chapter, we present a series of studies addressing this requirement.

4 Understanding mobility while preserving privacy

While the previous chapter presented results using individual mobility data, gathering such sets may raise privacy issues. Therefore, alternative approaches to sensing and studying human mobility are required. In the current chapter, we present and discuss results from studies that investigated the use of passive wireless measurements to assess mobility of a crowd. In the two studies presented, we perform these mobility assessments using wireless signals that were originally designed to serve a different purpose, namely Wi-Fi probe requests and Bluetooth object trackers. The major advantage of using the proposed approaches is that UIDs do not have to be permanently stored, however with an increased uncertainty when compared to individual measurements from a group of subjects.

Common solutions for research on crowd assessment use images (*e.g.*, photos or videos) which can imperil individuals privacy [15]. Alternatively, sensors in mobile wireless devices have enabled researchers to obtain large amounts of information about crowd behavior. This tracking ability, however, may be performed without the subject awareness, which forced regulators in many countries to enforce restrictive measures to its use (*e.g.*, [157, 158]). Therefore, crowd assessment systems must preserve their subjects privacy, while correctly inferring the requested metrics (*e.g.*, size, flow, speed).

In this chapter, we also discuss how the timing between sensing and reporting a *tag* in the Apple FindMy service can disclose information about a victim, including their taken trajectory and even possible end destination. Furthermore, we present a simple protocol with which data can be transmitted through a victims' iPhones, without them being aware.

To better understand the existing foundational work in crowd assessment, we present next a summary of the state of the art, including work that is relevant to the results and discussions presented later in this Chapter.

4.1 State of the art

In this section, we discuss a series of articles, organized by topics associated to results we present in the next sections. We present studies on crowd assessments as well as mobile technologies used for sensing, providing data for such studies.

4.1.1 Crowd estimates

Crowd estimates with passive Wi-Fi signal measurements have been extensively used for pedestrian monitoring. Examples include, public events (*e.g.*, music concerts) used in mobility simulations [159], wide public spaces (*e.g.*, universities) for flow characterization [159, 160, 161] and forecast models for users' trajectories based on movements in urban environments [162].

Furthermore, Wi-Fi *Management Frames* were previously used to estimate office occupancy [163], also on public buses [164], as well as graph-based studies of daily encounters and interactions between subjects [165, 166]. Likewise, Wi-Fi Received Signal Strength (RSS) [167, 168, 169] and Channel State Information (CSI) [170, 171] have shown solid results in indoor localization and subject counting. There are still, however, several open challenges on the integration and scalability of systems towards supporting a unified decision. As a recent example, the COVID-19 global pandemic fostered crowd monitoring research, when addressing problems such as assessing social distancing [172] as an effective measure to curb infections [40].

Alternative image-based methods include crowd density [173, 174], social distancing assessment [175], and crowd localization (*i.e.*, frame-by-frame tracking) [176]. However, even though these methods yield robust and accurate results, their utilization in real applications often raises privacy concerns [15].

4.1.2 Mobile device technologies used in crowd sensing

In this subsection, we exploit technologies available in certain wireless mobile devices for crowd sensing. Next, we review the literature on some of these technologies.

Wi-Fi Management Frames

On the data link layer, the IEEE 802.11 standard defines three frame types: *Data Frames*, responsible for carrying the data payload, *Control Frames*, allowing any device to manage access to the medium (*e.g.*, Request to Send (RTS) and Clear to Send (CTS) frames), and *Management Frames*, assisting Wi-Fi devices to find and connect to a nearby wireless network. From *Management Frames*, null frames, beacons, and probe requests/responses have been explored in the past for sensing the presence of mobile devices for carrying a device identifier while not containing any network traffic information (*e.g.*, browsing data), and for being continuously transmitted by devices.

Wi-Fi enabled devices continuously scan for nearby APs, collecting information (*e.g.*, Service Set Identifier (SSID), RSS and security configurations) relevant for the device to choose the best network available. This scanning procedure can be done either passively, by capturing APs' beacon frames, or actively, by issuing probe requests which should be followed by a probe response in case a desired AP was listening. Additionally, null frames are transmitted to report the power state of the issuing device.

In our work, we record the aforementioned frames as they are used to represent the presence of nearby devices. Note that, to mitigate individual tracking through unique identifiers in management frames, manufacturers implement randomization schemes that

periodically change the MAC address used. It is important to highlight that the use of Wi-Fi frames is not limited to studying crowds. As their identifiers or probing mechanisms may identify individuals devices [177], tracking phones or laptops over space and time is possible, posing yet another threat to mobile users. While Wi-Fi management frames have been extensively used for crowd sensing and subject tracking [15], their simplicity has forced phone manufacturers to reduce its impact on users' privacy. One of the adopted methods has been to randomize MAC addresses [178, 179]. However, there have been a series of proposed methods to overcome these imposed limitations, with relative success [180, 177].

As the randomization process used by manufacturers to hide a device's identity is not well understood, we discard all randomized addresses in order to minimize possible biases introduced by this approach. Additionally, we assume a one-to-one matching between devices and subjects, in could have led our reported results to suffered from inaccuracies due to people carrying multiple devices.

BLE trackers

A recent study by Weller *et al.*, evaluated different BLE trackers and their cloud services [181]. Their study revealed a series of security issues, including privacy risks with all products tested, although it did not include Apple's FindMy as no commercial product was available at the time. Focusing exclusively on the Apple service, Heinrich *et al.*, dissected how FindMy, Apple's object tracking system, components work [182]. Their study reverse engineered the protocol used by lost devices, finders and how owners can retrieve available location reports for their tags.

Apple ecosystem

Due to its popularity and extensive adoption, there is a growing body of research evaluating the security of various Apple services. For example, a study by Martin *et al.* [183] reveals how the Handoff service can compromise MAC address randomization and enable the re-identification of devices pertaining to a user. The Handoff service allows users to seamlessly share application's context and clipboard content across multiple devices (*e.g.*, between a MacBook and an iPad). On another example, Stute *et al.* [184] dissected the Apple Wireless Direct Link (AWDL) by reverse engineering its protocol and providing a complementary Wireshark plugin to this proprietary system. These examples highlight the relevance of studying such systems as they are used by multiple users and may not be carefully examined as their open source counterparts.

4.1.3 Covert communication

In addition to facilitating communication and device tracking, mobile phones and other technologies have also been used to covertly transmit data out of systems. Privacy and security research include a large number of studies that exploit systems to covertly exfiltrate data using a variety of methods [185]. For example, systems papers have used HDD or keyboard LED indicators [186], hidden signals in audio which is hard to tell

apart from music or existing background noise [187], as well as audio modulated signals using fans and HDDs [188].

4.2 Crowd Estimation using Wi-Fi Shadowing

In this study, we use measured Wi-Fi management frame signals to estimate crowd size while not compromising the privacy of those being monitored. This is achieved through measuring and correlating Wi-Fi RSS changes due to the presence of subjects in a monitored area (*i.e.*, shadowing), from signals sent by nearby fixed devices. As a result, relative crowd sizes can be estimated in a university building, without requiring any unique device identifier (UID).

Furthermore, our approach ensures that individual mobility patterns remain anonymous and cannot be traced back to specific individuals. By focusing on the changes in Wi-Fi RSS caused by the presence of subjects, rather than relying on UID-based tracking, we eliminate the privacy concerns associated with traditional crowd monitoring methods. Our experimental results demonstrate the effectiveness of this privacy-preserving approach, as we observe strong correlations between the number of mobile devices present in the building and the corresponding Wi-Fi RSS values from stationary devices.

4.2.1 Background

We now briefly discuss Wi-Fi *Management Frames*, along with our method to estimate crowd size from RSS changes. We validate our observations using a zero-th order approximation by counting the number of unique non-randomized MAC addresses in our set.

RSS-Based Estimates

To assess the density of a crowd while not having to continuously track all nearby *mobile* devices, we use the changes in RSS (*i.e.*, shadowing) from signals sent by *fixed* devices. That is, the crowd estimation is done through the wireless signal attenuation due to the physical presence and wireless activity of people in the observed area. In this context, we compare simple signal attenuation models which are the basis of our proposed approach. We note, however, that the estimation of the exact coefficients of the path loss models is not in the scope of our work.

Free Space Path Loss The first wireless path loss propagation model we discuss is the Free-space path loss (FSPL) Ω, which only considers the distance between sender and receiver along with the wavelength of the signal, and it is a powerlaw of the form $\Omega \propto d^{-2}$. The FSPL for a distance d and wavelength λ is estimated with Equation 4.1 [189]:

$$\Omega(d) = 20 \log_{10} \left(\frac{4\pi d}{\lambda} \right) \tag{4.1}$$

When considering Equation 4.1, for a same distance, we can expect a Wi-Fi signal at 2.4 GHz to have a lower path loss than at 5 GHz. Additionally, the FSPL yields an upper bound for RSS, if we assume a fixed infrastructure of senders (*i.e.*, fixed distance and transmission power w.r.t. to our receiver).

Log-Distance Path Loss A more robust alternative to FSPL is the Log-Distance Path Loss (LDPL) model, which describes the powerlaw attenuation with an additional log-normal variability. This random variability allow the model to better represent real world measurements as those have been shown to be influenced by such logarithmic components. The estimation of LDPL is given by Equation 4.2:

$$L_{\text{total}} = \Omega(d_0) + \gamma \log_{10}\left(\frac{d}{d_0}\right) + X_\sigma \tag{4.2}$$

where d_0 is a experimentally chosen *reference* distance (often 1 meter), and $\Omega(d_0)$ is the reference path loss at at the reference distance, d is the actual distance between sender and receiver, γ is the powerlaw exponent describing the signal attenuation, and X_σ is a log-normal random variable with centered at zero and standard deviation σ.

Signal attenuation γ The exponent γ is proportional to the complexity of the path, *i.e.*, it is defined by the properties of the obstacles between sender and receiver. Due to its intricacy, it is often only determined experimentally, and in this work, we assume γ will only vary because of people and their actions between a set of fixed senders and receivers in an observed space.

Absorption Cross Section and Wireless Interference A commonly used model for the effect of humans on electromagnetic waves is the Absorption Cross Section (ACS) σ_a [190], computed as the ratio $\sigma_a = \frac{P_{\text{abs}}}{S_{\text{inc}}}$ of absorbed power P_{abs} and incident power density S_{inc}. The value of σ_a is influenced by the direction, polarization and frequency of the transmitted signal, as well as the mass, cross sectional area and absorption rate of a person's body. We assume the ACS to be the same for all individuals, adding a constant effect to the total path loss. Similarly, we assume the interference from other wireless signals from people's devices (*e.g.*, Wi-Fi and Bluetooth) to be constant across subjects.

In this study, we estimate group sizes using the models described above for signal attenuation, as described next.

Fixed Sender Model

Empty building We define the Path Loss (PL) between a stationary sender *tx*, and a stationary sensor *rx* in an *empty* space as $\text{L}_{\text{empty}} = \mathcal{P}_{\text{tx}} - \mathcal{E}_{\text{rx}}$, where $\mathcal{P}_{\text{tx}}$ is the transmission power, $\mathcal{E}_{\text{rx}}$ the received power. Note that $\mathcal{E}_{\text{rx}}$ is the maximum RSS observed as there are no obstructions to the signal.

Occupied building Similarly, we define the PL when the observed area is occupied as $L_{empty} + L_{people} = \mathcal{P}_{tx} - \mathcal{Q}_{rx}$, where $\mathcal{Q}_{rx}$ describes the RSS when people are present. As a consequence, we can write $L_{people} = \mathcal{E}_{rx} - \mathcal{Q}_{rx}$, discarding the transmission power $\mathcal{P}_{tx}$ by assuming a stationary the sender.

Given all the above observations on path loss, we test (and validate) the hypothesis that changes in L_{people} are proportional to changes in the number of people in an observed area (*i.e.*, $PL/\langle PL\rangle \propto N_d$), where the expected value for PL is given by Equation 4.3:

$$PL = \int L_{people}\, f(L_{people})\, dL_{people} \tag{4.3}$$

where $f(L_{people})$ describes the density function of L_{people}. This loss caused by people along the path is proportional to the log-normal distribution from Equation 4.2.

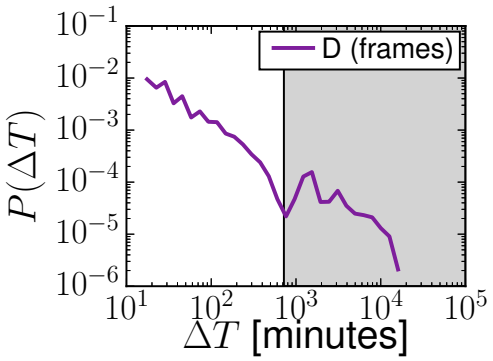
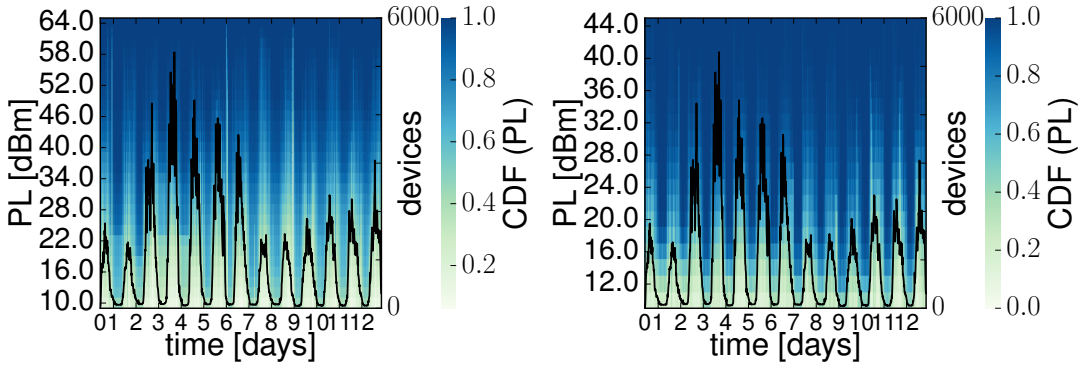


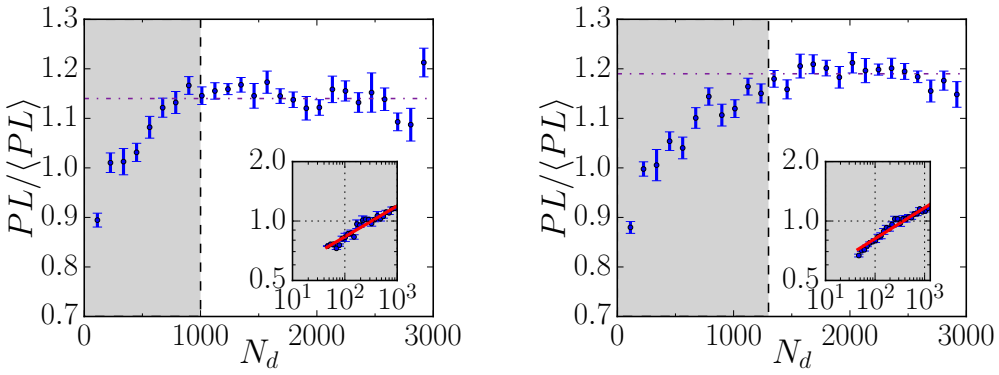
Figure 4.1: Stay duration with a bimodal distribution, with stationary devices in the gray shaded area.



(a) 2.4 GHz **(b)** 5 GHz

Figure 4.2: Average path loss (PL) and total mobile devices at different frequencies.

Finally, for an approximation of the changes in the number of individuals, we use changes in the availability of RSS of Wi-Fi Management Frames, which are periodically sent by wireless devices.



(a) 2.4 GHz **(b)** 5 GHz

Figure 4.3: Change in average path loss (PL/⟨PL⟩) with the total number of nearby devices N_d. An observed monotonic increase (shaded areas), followed by a saturation in the observed path loss. Insets show the shaded areas well approximated by $\text{PL}/\langle\text{PL}\rangle \sim N_d^{\alpha}$, with $\alpha = 0.156$ for both frequencies.

4.2.2 Fixed Infrastructure Evaluation

We now discuss the evaluation of our RSS-based approach to estimate crowd sizes, applied to a large university building. We first explain our experimental setup, then discuss our results where we observe a strong relationship between path loss and the number of nearby wireless devices.

4.2.2.1 Experimental setup

We installed 4 probe sensors in the Computer Science (CS) building of our university, near its main entrances, near a parking lot and near a cafeteria. This setup ran for 13 days, in February 2019, covering the lecture period for over 6000 students and 400 staff members who typically visited the building over a week. We will also refer to these location records as the *frames* data set.

Device Identifiers In our setting, each probe used 4 omni-directional antennae, listening to signals at the non-overlapping channels 6 and 11 (2.4 GHz), 36 and 44 (5 GHz). After parsing these packets, the devices' MAC addresses (*i.e.*, a device's UID) were classified as random and non-random. Then, frames from random addresses were discarded, while frames from non-random addresses had their UID hashed with a one-way hash function. This hashing was done in order to anonymize the true identity of these devices.

Our dataset From our measurements, we stored 28 million Wi-Fi frames, equally split between null frames and probe requests. These records include over 35,000 non-random (*i.e.*, global) MAC addresses, with a median presence of 1.1 days (25th-percentile: 51

minutes, 75th-percentile: 7 days, 99th-percentile: 13 days), and median time between records of 5 seconds (25th-percentile: 1 second, 75th-percentile: 16 seconds, 99th-percentile: 37 minutes). To reduce noise and eliminate passers-by, we discarded all records with RSS below -90dBm.

Next, we classify the observed devices from our set into stationary and mobile as later we compare the RSS from the former with counts from the latter.

4.2.2.2 Device Classification – stationary or mobile

We perform this classification by computing the PDF of sojourn times $P(\Delta T)$ of devices we observe. Similar to our approach in [6], we define a *stay* as a collection of consecutive records from a device, with a maximum interval of 15 minutes, where we also exclude any *stay* shorter than 15 minutes.

From the analysis of $P(\Delta T)$, we observe an inflection point at 12 hours, forming a bimodal distribution, depicted in Figure 4.1. With that observation, we classified devices as stationary if their *stay* duration was longer than 12 hours. This step resulted in our final list of mobile and stationary devices, to which we applied our proposed approach, and we present the results next.

4.2.2.3 Experimental results

Wi-Fi RSS varies along with device counts The mean PL, on both 2.4 and 5 GHz, from our identified stationary devices periodically changes as the number of *management frames* varies throughout the day, as shown in Figure 4.2. These panels display a color-coded Cumulative Density Function (CDF), along with the corresponding count of devices (N_d).

PL and device counts non-linear relationship Next, we correlate the mean PL and the inferred number of devices N_d using data from all four vantage points. For this analysis, we observe a monotonic relationship between the two variables, with a strong correlation up to 1,000 devices at 2.4 GHz and 1,300 at 5 GHz. This correlation is depicted in Figure 4.3, with insets showing an approximation for $\mathrm{PL}/\langle \mathrm{PL} \rangle \sim N_d^{\alpha}$, with $\alpha = 0.156$ in both frequencies.

PL and $\log(N_d)$ Pearson correlation is strong, and statistically significant Furthermore, PL and $\log(N_d)$ showed a very strong Pearson correlation $\rho_{2.4\ \mathrm{GHz}} = 0.74$ at 2.4 GHz and $\rho_{5\ \mathrm{GHz}} = 0.82$ at 5 GHz for the corresponding increasing intervals, also both statistically significant, *i.e.*, p-value $< 10^{-4}$. And as previously discussed, the correlations were not statistically significant on the corresponding non-increasing intervals, *i.e.*, p-value > 0.25.

The study collected a dataset of 28 million Wi-Fi frames and classified devices into stationary and mobile based on their sojourn times. The results showed that the Wi-Fi RSS varied with the number of devices, indicating a non-linear relationship. The mean path loss (PL) and the inferred number of devices exhibited a strong correlation, with a

statistically significant Pearson correlation coefficient. These findings demonstrated the feasibility of estimating crowd size while protecting privacy.

4.3 Crowd Estimation using BLE trackers

In this section, we discuss the results of re-purposing another system for crowd sensing. This time, we use reports for *lost* BLE devices, part of the Apple FindMy object tracking service. Similar to the Wi-Fi study, we perform real-world measurements of group size and people flow, validating our observations with assessments done using image recognition as well as management frames.

4.3.1 Background

BLE object trackers, or *tags*, are becoming increasingly popular and ubiquitous. This omnipresence allowed developers to build a network of tag owners, capable of reporting about nearby trackers without ever disclosing their identity. The benefits of this crowd-sourced effort are two-fold: users can swiftly find a lost object, and it creates possibilities for alternative functionalities that were not originally planned for. In this first study, we delve into one of these secondary uses built on top of Apple's FindMy service, namely crowd sensing. Next, we describe how this finder network service works.

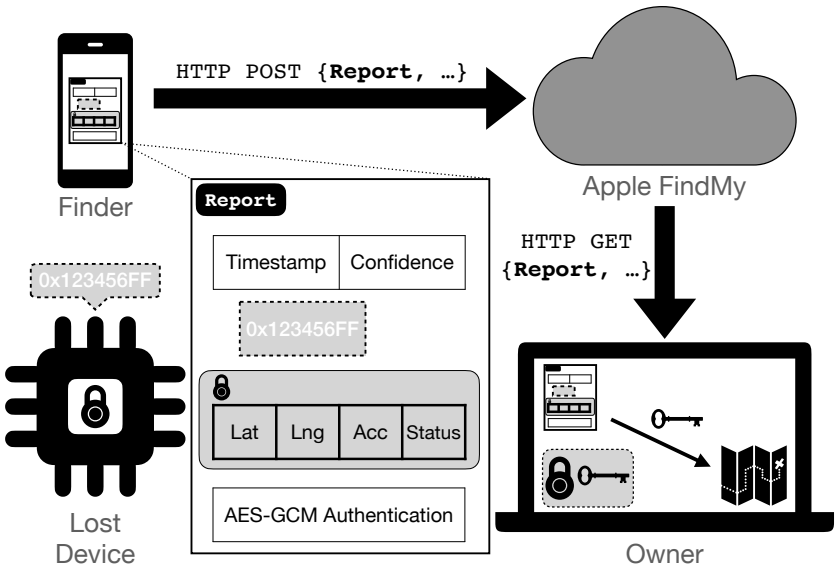


Figure 4.4: Delay in sensing and reporting a tag [1].

4.3.1.1 Finder network service

The Apple FindMy service was first released in 2019, allowing a device of a user who *explicitly* opt-in to be continuously tracked by other devices, in a crowd-sourced and

anonymous manner. As soon as a device is marked as *lost*, its owner will eventually receive reports containing the time and location where the device was (last) seen. Currently, this information can be seen through the FindMy app [191] on Apple devices.

FindMy reverse-engineering. The recent study by Heinrich *et al.* [182] unveiled the operation of the FindMy system, allowing any BLE enabled device to be used as a tag in Apple's finder network. These tags can be programmed to beacon about their presence at all times, which can then be captured and reported anonymously by *any* nearby finder, as depicted in Figure 4.4 which we describe next. As a first step, the owner of a tag must have an iCloud account, with which a pair of public-keys (e_k, d_k) is created through a series of API calls.

Tags beacons. To be tracked, a *tag* must broadcast BLE advertising packets using a crafted MAC address derived from e_k (see [182] for more details). Then, any nearby Apple *finder* device checks if a captured beacon contains a "matching" MAC address and payload, before anonymously reporting its presence to the iCloud service.

Finder reporting mechanism. Once a *finder* captures a valid beacon from a lost *tag*, it will temporarily store a report containing: (1) time t_c, (2) a confidence value c (similar to (4)), (3) e_k, (4) the encrypted location information using e_k, containing coordinates, horizontal accuracy and status, and (5) an AES-GCM authentication label for validity. Eventually, these reports are uploaded to iCloud using an HTTPS POST, where they are stored until an owner requests them. It is interesting to note that reports are *bundled* before being uploaded, and each record in a bundle receives an additional *time of arrival* (on the server side) timestamp, namely t_r.

Owner requesting reports mechanism. Using its set of e_k, an *owner* may query iCloud (HTTPS GET) for available location reports. If found, reports are then decrypted using the matching d_k.

BLE advertising. BLE advertising packets can carry up to 31 bytes of information, and can be continuously sent at intervals ranging from 20 ms to 10 s at three exclusive channels [192]. This variety leads to a stochastic reception of beacons, which is also influenced by the transmission power, distance between sender and receiver, and environmental conditions (*e.g.*, radio interference).

4.3.1.2 Experimental Setup

To further our understanding on how the sensing and reporting of *tags* in the FindMy network work, we conduct a series of controlled experiments. For that, we use microcontrollers (*i.e.*, ESP32) as our *tags* for supporting BLE and their low-power and easy programmability, allowing full control of its wireless chip. We show a sample of 10 location reports corresponding to a single tag in Table 4.1.

Table 4.1: Sample reports for a tag in the Apple FindMy.

confidence	capture time (t_c)	latitude	longitude	accuracy	report time (t_r)
2	1625672452	48.125760	11.590990	118.0	16256722471.748
2	1625672438	48.127524	11.593224	115.0	16256725090.865
3	1625672349	48.128720	11.593177	76.0	16256728238.983
1	1625672248	48.126633	11.600232	200.0	16256722694.255
1	1625672234	48.136451	11.570301	252.0	16256725643.866
1	1625672225	48.128819	11.593473	197.0	16256723496.669
1	1625672220	48.126010	11.601523	200.0	16256722484.134
1	1625672219	48.126592	11.600317	200.0	16256726498.278
1	1625672219	48.126010	11.601523	200.0	16256722484.134
1	1625672213	48.126697	11.605043	206.0	16256725625.544

Device settings determine when reports are uploaded. As the upload of reports is bundled, we analyzed the traffic between iCloud and jailbroken iPhones[1] to study the reporting behavior for our phones using different settings (*e.g.*, cellular or Wi-Fi, and on battery or power supply). For each of these settings, we placed a continuously beaconing *tag* next to our phones for 72 hours. From these measurements, we observe a clear distinction between a phone's connectivity modes, with a median reporting time of 15 minutes on Wi-Fi and 3 hours on cellular, as depicted in Figure 4.5. Additionally, when Low Data Mode was enabled, our iPhones submitted reports with a median time of 36 minutes. These numbers lead up to our first observation that phone settings significantly influence the difference between t_c and t_r, *i.e.*, the expected time reports are bundled.

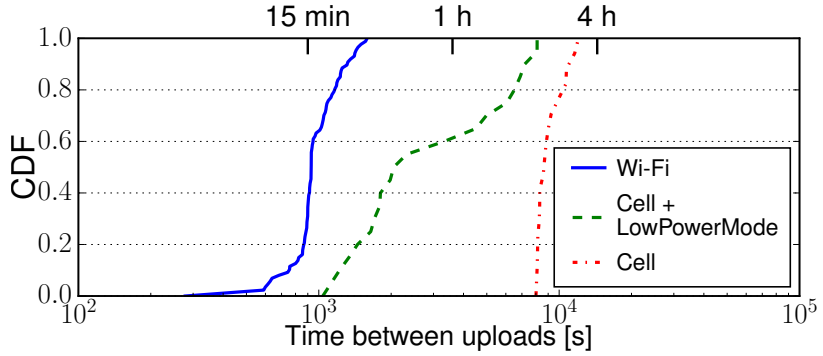


Figure 4.5: Delay in sensing and reporting a Bluetooth tag.

Most reports are uploaded within 15 minutes – in the wild. To complement our controlled measurements, we collected location reports from our own *tags* for 24 hours in a crowded public space, over multiple days from July to September 2021. The difference between contact time t_c and reported time t_r showed a strong mode around 15 minutes

[1] iPhones 7 and 8 on iOS14.6, through an HTTP proxy

(median of 13.15 minutes), and with 95% of its values between 6 seconds and 8 hours as depicted in Figure 4.6. This highlights the variety in reporting behavior, which we will exploit for our crowd assessments.

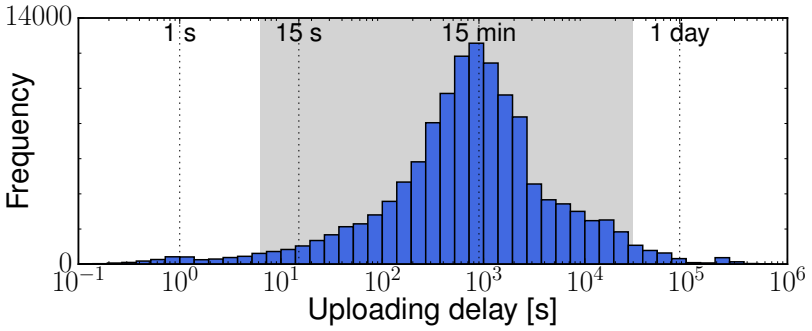


Figure 4.6: Distribution of the delay in sensing and uploading.

Bundle uploading time uniquely identifies a finder. Report bundles are appended with their receiving time t_r with milliseconds precision, allowing us to *uniquely* identify a *finder*. That is, even if two phones sense a beacon from a tag at the same time, the reporting to iCloud is unlikely to reach the Apple servers within the same millisecond. This identification is made stronger when bundles include several reports, with wider ranging contact times t_c. It is worthwhile noting that bundles may include up to 255 reports, with up to 4 reports per *tag*.

Quicker BLE advertising may lead to no availability of reports. As previously discussed, the sensing of BLE advertisements is stochastic, and shorter advertising intervals will lead to a higher success rate in their detection. However, our experiments showed that at 20 ms or lower, the FindMy service discarded all location reports, possibly for fairness. Although we were not able to precisely pick the best interval, our experiments showed that at 1022.5 ms, we achieved the lowest miss rate for detected beacons. With that observation, we used this interval for all our future measurements, at maximum TX power (+9dBm).

4.3.2 Crowd Monitoring using tags

We now assess the use of *tags* from the Apple FindMy service in two problems in crowd monitoring: (1) using a single *tag*, we estimate crowd density in an area, and (2) using multiple *tags* we gauge crowd flow over a path. For their evaluation, we compare (1) against a state-of-the-art deep learning method, whereas we compare (2) to our previously discussed Wi-Fi management frames approach.

4.3.2.1 Crowd density – Using a single tag

To evaluate the use of *tags* for crowd size estimates, we run 8 measurements, each for 3 hours, from July to September 2021. These measurements were conducted in a popular public square, Marienplatz in Munich, Germany, that it is often visited by those coming to shops, restaurants, and nearby transit stations. In this experiment, we used a single ESP32 as a *tag*, advertising at a constant interval of 1 second and maximum TX power (+9 dBm). The numbers obtained from this approach were compared to those obtained from an image-based approach, that we explain next.

4.3.2.2 Image-based crowd count

Image recognition methods applied to crowd estimates yield accurate and robust results, even with inexpensive hardware [15]; solutions based on CNNs have lately produced unrivalled results for computer vision. Despite its superiority, it may still raise privacy concerns for the use of images which lead to regulatory legislation in various countries. To produce a comparative metric, we applied a CNN model to images from a webcam[2][3], which publicly publishes 2048x1536 images every 5 seconds. For the recognition task, we used the `Mask R-CNN` [193] model, which performs image segmentation for multiple class instances in parallel. This segmentation is important to better separate individuals in a frame, as they tend to stand and move in large groups, even when they overlap in an image.

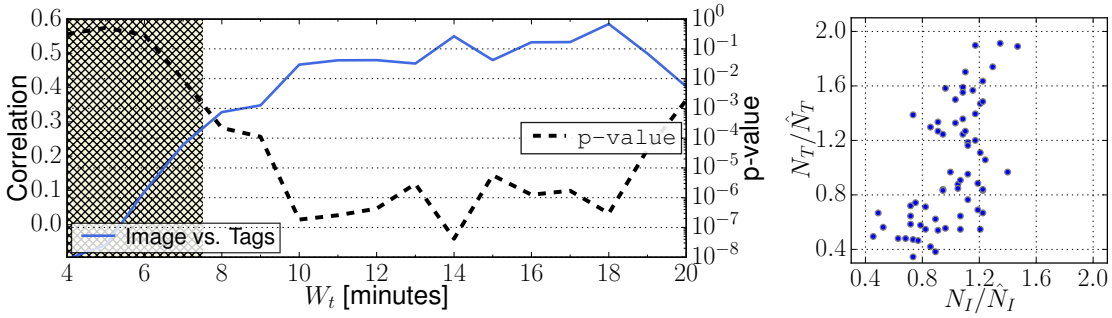
4.3.2.3 Results of crowd size estimates

For this analysis, we correlate the number of individuals recognized by the image-based approach against the counts from *tags*. For the latter, each device is identified by the number of unique receive timestamps t_r over any considered interval. In the image-based method, a CNN model called `Mask R-CNN` was applied to images captured by a webcam. The `Mask R-CNN` model performs image segmentation for multiple class instances, allowing for better separation of individuals in a frame, even when they overlap. In this way, we compare the image-based approach with the *tags* method to determine the accuracy of the former in estimating the number of individuals.

To choose the best window to aggregate location reports, we correlate the two estimates (*i.e.*, *tags* and images) for different time bin sizes (W_t). In this analysis, we observe that only window sizes above 8 minutes yield statistically significant results (*i.e.*, we adopt p-value < 0.001 as our accepted confidence interval), and the correlation reaches its peak value at 18 minutes (at 0.58), as depicted in Figure 4.7a, with its corresponding correlation plot on Figure 4.7b. The higher values of correlation around 15 minutes are likely explained by the reporting behavior discussed previously, yielding crowd size estimates sizes with a tolerable time lag.

[2] `https://kaufhaus.ludwigbeck.de/service/webcam`
[3] No explicit authorization was obtained for the use of these images, which are publicly accessible.

As a result, estimates of crowd size could be best estimated using a window of 18 minutes. This could then be calibrated with an actual counting during limited intervals, and using a model based of the report counts to then estimate sizes at any other time.



(a) Correlation (N_T and N_I) and bin sizes (W_t)

(b) $W_t = 18$ mins

Figure 4.7: Relationship numbers from tags (N_T) and from images (N_I), for different bin sizes (W_t). ($\hat{N}_*$: mean N_*)

4.3.3 Crowd flow – Using multiple tags

We now evaluate the use of *tags* to estimate moving and waiting times for a crowds in a public space. In this case, we compare our results to measurements done using Wi-Fi management frames, done from two vantage points in the center of Munich.

4.3.3.1 Wi-Fi Management Frames

As previously discussed in this thesis, this approach produces acceptable results by tracking devices using their unique MAC address. In this experiment, we discard all Wi-Fi frames containing a locally managed address (*i.e.*, random[4]) in order to reduce unwanted noise, which accounted for 26% of the total records being removed. For these measurements, we used two RaspberryPi 3, with two external antennae, hopping between channels 1, 6 and 11.

4.3.3.2 Results of crowd flow estimates

To study the flow of a crowd, we compute the interval devices take to traverse the space between two vantage points, as measured using our *tags* approach and using Wi-Fi signals. Finally, the evaluation is done by comparing the distribution of these time intervals between the two approaches. Our measuring points were 176 meters apart, located in the city center of Munich, at a busy pedestrian only area near shops and transit stations. We made 3 measurements, which lasted for 2 hours each in the early days of September 2021.

[4]Wi-Fi enabled devices randomize their MAC identifiers when sending probe requests to avoid being tracked (see 4.1).

As previously discussed, we identify a device in the *tags* measurements by their reporting time t_r (see Table 4.1 for example), from a bundle containing consecutive records from both vantage points. Similarly, from the Wi-Fi measurements, records from a non-random MAC address captured at both fixed points yields a valid interval.

We first note that the two distributions (*i.e.*, from *tags* and Wi-Fi) show a satisfactory similarity, with a strong mode at 2 minutes, as depicted in Figure 4.8. Additionally, we decompose both distributions into their log-normal components, with an unsupervised learning model, namely Gaussian-Mixture Model. This popular method clusters data into a given number of Gaussians (n_G).

To choose n_G, we applied the BIC method [194], which yields a score for how well each model (*i.e.*, the labeled data) explains variance in the input data. This approach suggests 3 to be the best value for n_G, which when applied to the distributions discussed above, yields the right panel on Figure 4.8. Note that we discard one of the clusters, which contained very short intervals.

We interpret the remaining curves as the walking times and walking + waiting times. Both sets presented similar average walking times, with 2.41±0.04 minutes for *tags* and 2.28±0.04 minutes for Wi-Fi frames. These times correspond to a walking speed of 4.5 km/h, which is in line with existing research [195]. Furthermore, given this walking speed estimate, we observe similar average waiting times of 19.33±1.44 minutes on *tags* and 20.54±0.69 on Wi-Fi frames, *i.e.*, the possible time people spent at shops along the way. This allows us to estimate moving and stop times using the location reports for our tags, and given the lack of device identifier, this method cannot expose the actual identify of any nearby subject.

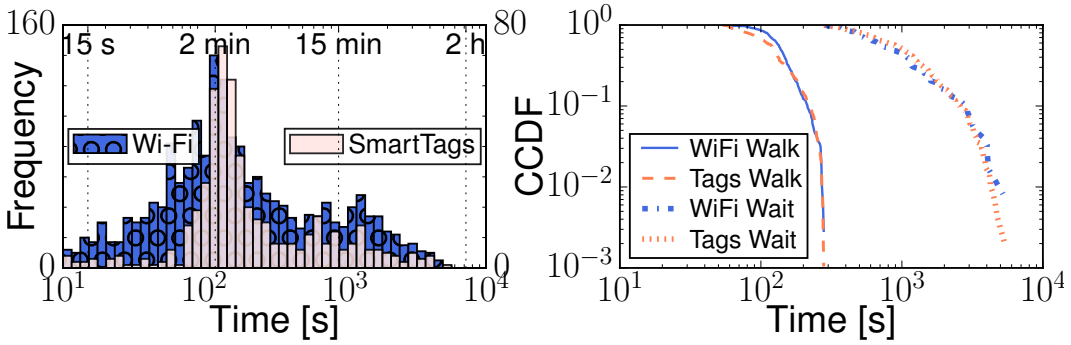


Figure 4.8: Crod Flow [Left] Time between vantage points. [Right] Estimated walking and waiting times between vantage points.

4.4 Privacy leakage on the Apple FindMy service

In this Section we expand on the analyses of *tags*, presented in Chapter 4. This time, we unveil a series of possible attacks iPhone users could be subject to, along with proof-of-concept examples and a discussion on possible mitigation strategies. We present

these attacks in two parts: deliberate tracking and side-channel communication. For background on the working of this finder service, please refer to Section 4.3.

Subject Tracking Beyond the sensing capabilities of the Apple FindMy (see § 4.3), we identify possible risks to users privacy due to information contained in the timing information of location reports.

Side channel Additionally, we show how a combination of *tags* can be used to establish a covert data channel, able to send data through a victim's iPhone at low bit-rate, without a victim noticing it.

4.4.1 Deliberate Tracking

We now introduce a series of PoC evaluations of location information that could currently leak from iPhones through the Apple FindMy service. These threats that we discuss rely on *timing attacks*, in which the difference between sensing and reporting a *tag* may give an attacker sensitive information about a victim. Furthermore, the attacks we discuss were performed on our own devices, where we had control over the iPhones' settings and times when they were connected to cellular or Wi-Fi.

We present two possible attacks of location information leakage, that rely on location reports being bundled (see § 4.3): A) *Final destination inference*, where typical phone settings and the difference between sensing (t_c) and uploading a report (t_r) can give valuable information to an attacker about a victim's whereabouts. B) *Partial path reconstruction*, where location points covered by a phone's trajectory may be inferred.

4.4.1.1 Remote Destination Inference – Using a single tag

This PoC is based on the time information from location reports, namely, the time between observing a nearby *tag* (t_c) and sending a bundle of reports to iCloud (t_r). Additionally, we note that if supported by the *tag* hardware, the TX power can be modulated to ensure *only* a victim's iPhone is affected.

Threat model In order to obtain relevant information about the whereabouts of a victim, a malicious actor conducts a timing attack with a single *tag*. This actor is aware of the victim's mostly visited places, and a victim's iPhone only uses Wi-Fi when visiting these places, while on cellular at other time. During a brief encounter with the victim, the attacker "pins" their iPhone by transmitting a series of beacons with a crafted *tag*. The pinned phone, while on the move, will be connected to the cellular network, storing reports until reaching the victim's final destination. At the victim's final destination, their phone will connect to Wi-Fi, immediately uploading the existing location reports. Finally, by knowing when the contact happened and the time it took for the reports to be uploaded, the malicious actor can infer the most likely final location by comparing possible travel times. While this attack happens, the victim is ignorant

that information is being stolen and only disabling the FindMy service, or Bluetooth, can currently mitigate this threat.

4.4.1.2 PoC

To test this attack, we used one iPhone 12 (running iOS 15) and one crafted *tag*, with TX power modulated at -6 dBm to limit the reachability of each beacon to only our phone. We then powered up the *tag* close to the iPhone for 60 seconds, then traveled 18.5 km to a location, where we enabled Wi-Fi. As a results, we observed that the total time on the move was 29 minutes, and $t_r - t_c$ was 35 minutes. That is, the time to report closely matched our travel time, with a difference due to probably the phone taking a few minutes to connect to Wi-Fi at our final destination. Note that in this proof-of-concept experiment, we do not collect enough samples to cover all possible cases, when for example, other phones are present.

4.4.1.3 Path reconstruction – Using multiple tags

Given the bundling of reports (see § 4.3), intentionally positioned *lost* tags can form a sequence of "breadcrumbs" which can then disclose the path followed by a phone. Similarly to the previous example, TX power can be modulated to ensure shorter coverage from each tag. As a proof of concept, we conduct one experiment.

Threat model A malicious actor places crafted *tags* along possible paths their victim's might move through. This attack requires a victim's phone to be connected to cellular at all times while moving through this monitored area, and it may connect to Wi-Fi after leaving it. As the victim's iPhone moves through this watched area, it will sense the different *tags* along the way and bundle their reports. Once this bundle is eventually uploaded, the time and known location of the *tags* allows the attacker to reconstruct the path taken. Note that, the t_r of a bundle is unique, allowing consecutive reports to be grouped (see § 4.3). Once again, while the attack is happening, the victim is not aware it is happening, and only disabling FindMy or Bluetooth can currently mitigate this attack.

4.4.1.4 PoC

In this case, we placed 6 different *tags* in pairs, 150 meters apart, and we conducted experiments with 3 different iPhones (7 and 8 on iOS 14.8 and 12 on iOS 15). We stood next to a pair of *tags* for 5 minutes, then moved to a next location (in approximately 2 minutes), with Wi-Fi disabled in all phones. Once all three spots were visited, we moved to another distant location where Wi-Fi was finally enabled. The wireless configuration was set to ensure reports were not sent before all stops were covered.

Based on the logged contact times (t_c), we were able to infer the path taken by all three phones, along with their stops, as depicted in Figure 4.9. We also noted that reports were uploaded within 5 minutes of the iPhones connecting to Wi-Fi, enabling a malicious actor to efficiently perform this attack, unless partial bundles of reports are before all

stops are covered. It is important to note that with the wide availability of public and private Wi-Fi networks, *e.g.*, `eduroam` and network operators offloading services, such attacks could be made even easier. While such networks provide continuous Internet connectivity they may also allow a near-real-time tracking of subjects.

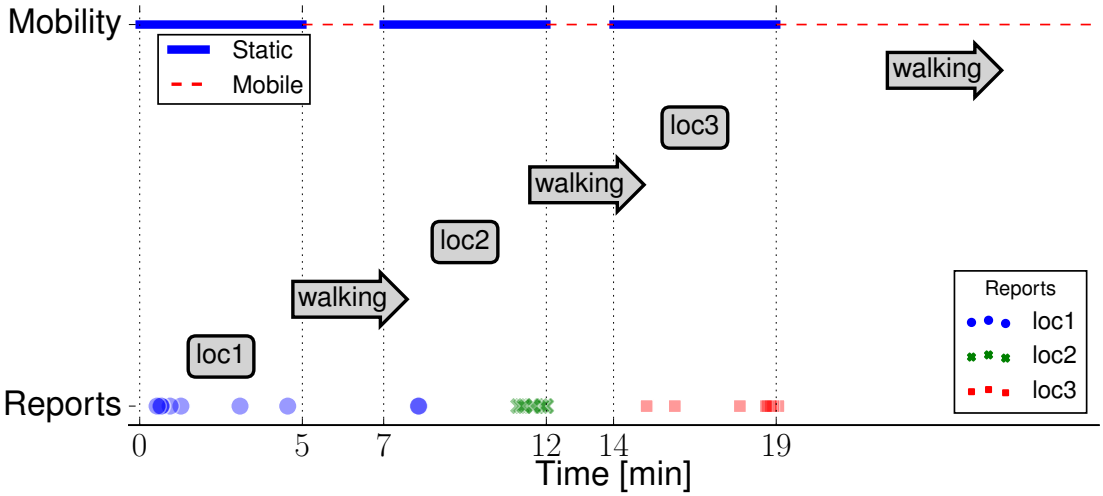


Figure 4.9: Path reconstruction.

4.4.1.5 Mitigation

Until Apple chooses to change (or fix) these behaviors, which we discuss below, iPhone owners can mitigate risks of having location information leaked by disabling the FindMy service or Bluetooth. However, note that this may reduce the existing functionalities of the phone which could be prevent if Apple addresses the issue. These possible fixes include: (1) randomizing the time it takes to upload a report, making it agnostic of the connectivity available, or making it a function of the distance traveled, or (2) reduce the granularity of t_r, or remove it, relying on t_c to tell whether or not a report is stale. These suggestions should not affect the functionalities of FindMy, but could significantly improve a user's privacy protection.

4.4.2 TagComm – Covert Channel Using BLE Trackers

Complementary to the leakage of location information in the Apple FindMy service, we now present a possible side-channel attack on this tracker service. Our unicast protocol, named *TagComm*, encodes data as sequence of different *tag* beacons, which when sensed by a nearby iPhone, can be covertly transmitted to an attacker through the iCloud reporting system. Figure 4.10 graphically presents *TagComm*: a crafted *tag* switches between a set of IDs, that are chosen to encode a message using a pre-determined alphabet, then a *finder* senses these IDs which are later reported, allowing the *owner* (*i.e.*, the attacker) to decode the desired message.

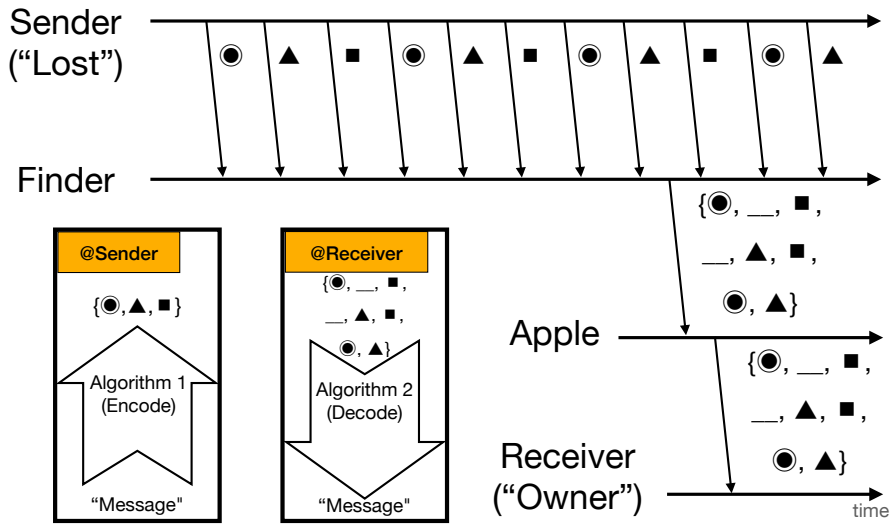


Figure 4.10: *TagComm* protocol example, encoding a message as a sequence of tag IDs, silently and securely transmitted by a *finder*.

A malicious actor could, for example, use our system to exfiltrate data from an air gapped system (cf. [185]). Additionally, as the Apple FindMy reporting system is end-to-end encrypted, neither Apple nor a victim's iPhone (*i.e.*, a *finder*) would be able to decode any message or note that the attack is taking place. Only the *owner* of the transmitted IDs would be capable to decode the exfiltrated message.

TagComm encodes information with a series of *tag* IDs, and their permutations. This approach allows `N` IDs to encode $\lfloor \log_2(N!) \rfloor$ bits of information. Furthermore, we add a set of additional bits for increase robustness, including header and parity bits.

4.4.2.1 Encoding a message

To maximize bandwidth and support message validity checks, we encode the data to be transmitted using a permutation of `N` *tag* IDs. Note that the understanding of how *tags* are sensed and how location reports are built and uploaded (see § 4.3) limit the achievable bitrate of this approach.

Input to sequence of symbols Considering a word `W` to be transmitted, and an encoding alphabet of symbols of size `N`, we split the interval (0, `N!`) iteratively to recover the sequence of symbols to be used. Note that this requires `W` < `N!` (as ❶). For example, for `N`=5 and `W`=42, we first sort the set of encoding symbols, *i.e.*, a<b<c<d<e (❷). Next, we divide the interval 5! into 5 blocks of equal size, as in Figure 4.12a. As 42 is in the second block, *b* is set as the first encoding symbol. This is repeated until all symbols have been assigned, in this case, outputting *bdeca*. This approach allows us to encode $\log_2(N!)$ bits using `N` tags (*i.e.*, symbols). Algorithm 1 systematically describes these steps.

Algorithm 1: Encoding input word into sequences

```
   Input: S,w; /* Symbols and word encoded                          */
   Output: E;                                  /* Encoded sequence */
1  L ← length(S) ;                                /* Length of S */
2  assert(L! >= w) ;                                   /* From ❶ */
3  E ← [ceil(w/(L − 1)!)] ;                      /* One item list */
4  for idx ∈ range(L − 1, 0, −1) do                    /* From ❷ */
5  |  w −= (E[−1] − 1) ∗ idx! ;                    /*E[−1]: last */
6  |  e = ceil(w/(idx − 1)!);
7  |  E.append(e);
```

Defining N=16 Given ❶, we define the code efficiency of our protocol as $\lfloor \log_2 N! \rfloor / \lceil N \log_2 N \rceil$ (as ❷). That is, the maximum encoded bits by the minimum bits for all symbols N. From these observations, Figure 4.11 shows the change in ❷ for different values of N up to 20 tags[5], with its highest efficiency at N=16, that we then we use for our protocol. This setup yields 44 bits in total, which use we describe next.

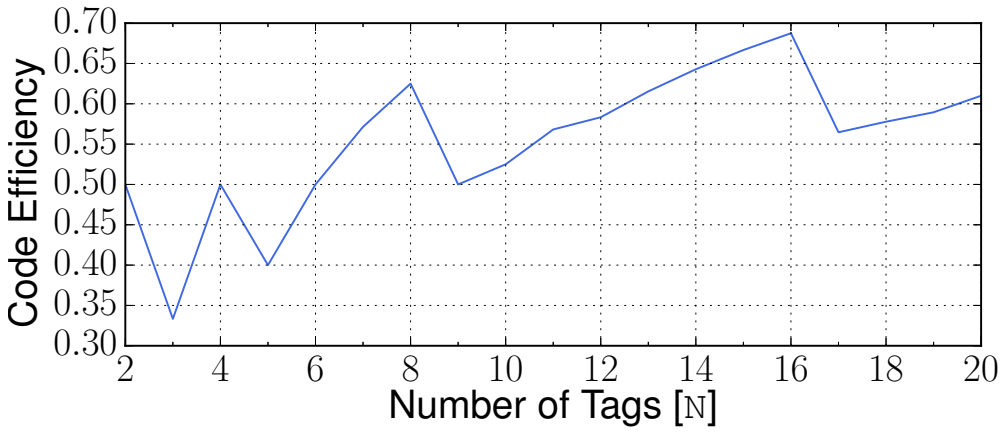


Figure 4.11: Code efficiency given the number of symbols (tag IDs) being used to endcode a message, with its maximum at 16.

Frame and supporting bits In order to allow the receiver to verify the integrity of the message being transmitted, we define a series of special bits for our protocol frame, depicted in Figure 4.12b. Our header, consisting of three bits, differentiates among the following message types: STX (0b000) the start of transmission, F0 and F1 alternating even and odd frames (0b010 and 0b100, respectively), and EOT end of transmission (0b110). By using these bits, a receiver can deterministically identify the start, the end

[5]Largest number of permutations that fits in 64 bits.

and intermediate frames during a transmission. Additionally, a parity bit is included to allow for an integrity check of each frame. The remaining 40 bits are our payload.

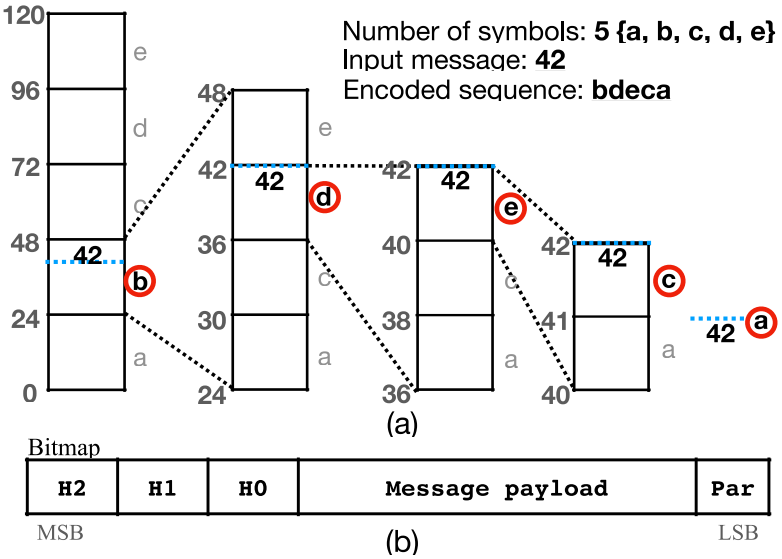


Figure 4.12: Protocol definitions. (a) Encoding a value as a sequence of symbols. (b) Frame bitmap.

Transmission integrity To successfully send a message `W` using TagComm, a sender has to chunk it into N_w words `w` of size 40 bits. An `STX` frame (*i.e.*, header set to `0b000`) starts the transmission, also containing the total number of words N_w to be expected. Next, each `w` is encoded and transmitted in alternating frame types `F0` and `F1`. Ultimately, an `EOT` frame signals the successful transmission of the message. Note that, for every `w` all symbols must be transmitted and successfully received.

4.4.2.2 Decoding a message

Once the location reports have been received by an *owner*, the transmitted message can be decoded. This is done by reversing the steps described above for encoding.

Tags alignment The first step to decode an incoming message is to align each symbol (*i.e.*, *tag* ID) into slots of a predetermined duration t_{tag}.

Frames alignment As the duration of each frame is also predetermined, the next step is to align the received frames. With that, each frame type can be identified and accordingly aligned, *i.e.*, `STX`, `F0`, `F1`, ..., `EOT` (as ❸), allowing us to start the decoding of the message.

Decoding frames With the frames aligned, we can decode the message. This is done by adding up the partial values resulted from dividing the interval `N!` and knowing the

predetermined ordering of symbols (*i.e.*, ❷), as in Algorithm 2. Finally, in the next step we can then check for errors using the parity bit.

Algorithm 2: Decoding sequences into words

```
   Input: S, E;                                  /* Symbols and encoded seq.  */
   Output: w;                                    /* Decoded word */
1  L ← length(S) ;                               /* Length of S */
2  P ← [ ] ;                                     /* Empty list */
3  for i, e ∈ enumerate(E) do
4  |  idx ← S.index(c) ;                        /* Symbol e index */
5  |  P.append((L − 1 − i) ∗ idx);               /* Partial sum */
6  |  S.pop(idx)
7  w ← sum(P) ;                                  /* Add up all partials */
```

Error identification After decoding all frames, we check for errors with the parity bit. Additionally, the sequence of frames (*i.e.*, ❸), along with parity bits, enables us to reconstruct the original message when a *single* symbol is not received. This is done by first aligning the received frames, then correcting the error with the corresponding parity bit of the faulty frame.

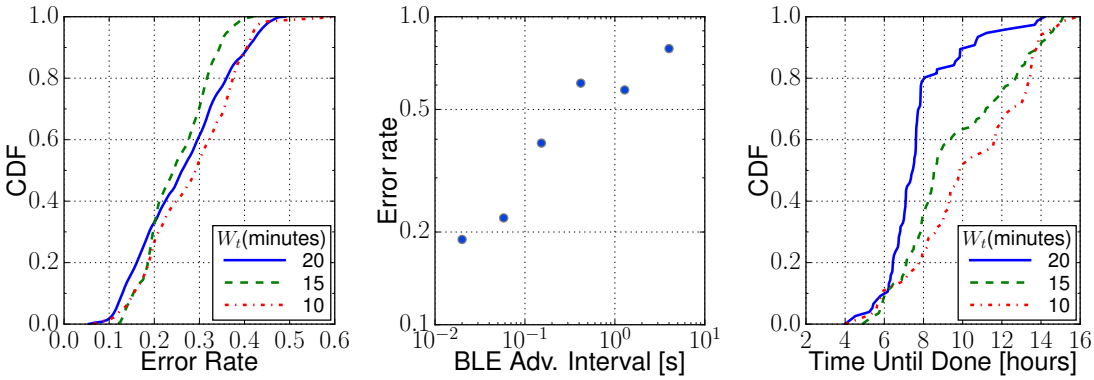
Final message reconstruction With all frames decoded, the number of transmitted frames N_w can be verified, and the final message can then be reconstructed.

4.4.2.3 TagComm Experiment

We now describe the tests and results obtained from our TagComm system.

Setup and results To test our system, we transmitted 10 random words of 40 bits, using a single ESP32 as a *tag* running our TagComm implementation. As a *finder*, we used a an iPhone 12 (running iOS 15) which was always on Wi-Fi. As discussed in § 4.3, these settings represent a best-case scenario in which reports are expected every 15 minutes. With this setup, we successfully received all transmitted words. Furthermore, we tested the boundary conditions for TagComm, such as word duration and BLE advertisement interval.

For these additional analyses, we define error rate as slots where no location reports were received. That is, for a tag slot (*e.g.*, 30 seconds), the corresponding error rate of a transmission is the number of empty slots a report was expected. Additionally, the Time Until Done (TUD) is the least amount of time that was required to decode the entire information.



(a) (b) (c)

Figure 4.13: Error rate and TUD for different settings. (a) CDF of error rate and different frame duration (W_t). (b) Error rate and BLE advertisement intervals. (c) CDF of TUD and W_t.

Frame duration shows little impact on error rate From our tests, we observe a 26.7% error rate when using 20 minutes frames, 24.73% for 15 minutes, and 28.82% for 10 minutes on average, as shown in Figure 4.13a. This suggests that increasing the number of times a *tag* is transmitted within a frame window does not significantly affect its probability of being detected, assuming an iPhone is around.

Larger BLE advertisement intervals increase error rate As shown in Figure 4.13b, the error rate monotonically increases with larger BLE advertising intervals. This is likely due to the stochastic properties between these intervals and the detection of beacons [192]. That is, as a Bluetooth device is not continuously listening on all channels, if we reduce the interval between transmitting new beacons, we reduce the probability each transmission is sensed by a receiving device.

Shorter frame duration increases TUD Using frame windows of 10, 15 and 20 minutes, we tested the TUD of each configuration for 72 hours, finally we computed the expected TUD using 100 random starting points. With these tests, we see 7.73±0.22 hours with 20 minutes, 9.57±0.30 hours with 15 minutes, and 10.36±0.36 hours with 10 minutes for TUD, which are also depicted on Figure 4.13c. It is important to highlight that, in our tests, two iPhones connected to the same iCloud account when used a *finders* did not produce a statistically significant improvement in TUD.

In summary, the experiment demonstrated successful data transmission and reception. The error rates were affected by frame duration and BLE advertisement intervals. Increasing transmission frequency within a frame window had little impact on detection probability. Shorter frame durations increased the decoding time, while multiple iPhones connected to the same iCloud account did not significantly enhance it. These findings provide insights for optimizing the system's performance and efficiency.

4.4.2.4 Mitigation

As previously discussed for the leakage of location information through the Apple FindMy service, currently, users can only mitigate this threat by either disabling that service or Bluetooth altogether, impairing the device's functionality. Regarding the system, Apple can limit the number of concurrent reports that can be sent by a *finder*, as well as reduce the overall number of reports available for a lost *tag* or reduce the granularity of the timestamps contained in each report. During our experiments in 2021, we noted that iOS never notified our phones about the presence of a nearby crafted *tag*, which could also help as it is currently done for AirTags. That is, if an AirTag is detected near an iPhone for an extended period of time, the phone user will be notified that a tracker is likely follow them. The same behavior is not implemented for the crafted tags we used.

We believe that the possible attacks discussed here are only of low risk for most iPhone users given its requirements and the sensitivity of the information leaked. Mobile users who deem such risks high can disable the FindMy service reducing their exposure at the expense of being able to track the location of their own devices remotely.

4.5 Discussion

This chapter demonstrates our findings on studying human mobility through aggregated data from multiple mobile users, as opposed to tracking individuals. We achieved this by detecting movement-related signals from mobile phones and aggregating them at the time of collection. Our investigations also exposed the potential leakage of private and sensitive data without users' awareness.

As such, we address the final two research questions, **RQ2** and **RQ3**, regarding minimizing privacy exposure while studying mobility and whether systems designed with "privacy by design" may still unintentionally leak information.

Our studies using Wi-Fi signals and Bluetooth trackers data demonstrate the feasibility and limits for studying crowds. While these results do not address the same problems covered on chapter 3, they expand our understanding of related aspects on human mobility, namely occupancy, density and flow.

Additionally, the study of these systems revealed how location information can leak and be exploited by actors with malicious intentions. That is, while the availability of services associated to mobile devices continues to grow, that may also unwittingly increase the surface of attack and exposure of users' private data.

5 Conclusion & Outlook

In this thesis we studied how mobile devices can be used to sense different aspects of human behavior as well as how they may influence these actions. That is, looking through GPS data or Wi-Fi associations to model mobility we study how browsing and online gaming may vary and even influence back mobility. We quantified these influences and explained them through mathematical models that can be reused and compared with newer observations. Furthermore, we discussed results of how new technologies and the omnipresence of devices may inadvertently expose individual privacy, and how behavioral studies on mobility can be conducted using aggregate measurements. These privacy risks raise important questions regarding the possibility of conducting such studies while minimizing exposure, also presenting clear ethical concerns related to the research being done. Our results highlight the importance of better understanding how people move and the consequences of these actions in their contacts and their wireless network utilization. We proposed robust models built from real world data and simulations that capture and describe fundamental aspects of human nature when mobile users are studied individually or as crowds. Furthermore, our work highlights potential cybersecurity risks wireless-based services may expose their users. That is, the introduction and wide spread use of mobile connected devices can be used to study human mobility, with clear benefits, but this should be done with privacy and ethics taking a central role in human behavior research.

In the first part of our results, we focused on individuals, and we used data from their mobile device usage to study mobility. We showed how different device types are used differently to browse the Internet and how that usage correlates with mobility, including their predictability. The quantification of these differences may help guide the implementation of new network protocols as well as user application interfaces, helping deliver the right content given a user mobile state. Similarly, our study on the interplay between online gaming and mobility shows how applications with a strong mobility component can significantly alter human behavior, and how this impact may be affected by the battery consumption of the game. These observations should foster further research on (1) mobile network protocols, improving connectivity, speed and latency given future actions of a mobile user; (2) on the different use and preference for information depending on the device type used and whether the user is on the move or not, also considering the changes in usage pattern over the years; and (3) on application design patterns that best suit applications with a strong mobility component, such as games, maps, and social media.

Still on the analysis of mobility through individual phone data, we studied how the duration of stays and the resulting social network of contacts can be used to estimate risks during an epidemic spread. The results of these studies could be used in future

epidemics, better guiding policy makers and sanitary guidelines. This could be done by making informed predictions on the time someone will spend at a location, as well as notifying people that could be at an elevated risk of infection based on their neighboring structure rather than only direct contact. We believe these observations could also have implications beyond infectious diseases spread, such as opportunistic networks, (mis-)information dissemination, and urban planning.

In the second part of our results, we kept our focus on data from mobile devices, but for the study of crowds instead. This shift allowed us to study different aspects of mobility, while attempting to preserve users privacy. Using Wi-Fi probes we showed how even the change in signal strength measured by several probes can correlate with the presence of people. This ensures privacy while allowing us to build estimates of the number of people in a confined and predetermined space. A challenge that remains here is the use of a similar setup in unknown environments and where the presence of individuals changes constantly. Through Bluetooth, a similar technology designed to short range communication, we built a system capable of estimating crowd size and flow on top an existing infrastructure by Apple. The FindMy service, capable of tracking lost devices and BLE tags, provides us with enough information to do crowd assessment with minimum privacy guarantees. That is, location reports for lost BLE trackers do not carry identifiers from finder devices. We have shown, however, that with the information currently provided with each location report (and some additional external information) this assumption no longer holds. We were able to show how individuals may be tracked and even have information relayed through this system unknowingly. These results highlight the potential risks the introduction of new technologies may bring to mobile users. While these risks should not prevent innovation, changes should be admitted with caution as they are likely to also introduce new, and unwanted risks to users privacy. We hope these observations will foster two major effects: (1) further research on crowd studies, including different data sources as well as the use of these insights for policies and commercial use, and (2) the privacy and cybersecurity risks brought by systems that were designed to preserve users identity but could be easily compromised.

Privacy and ethics were central components of this thesis and the scientific papers presented. When personal data from mobile users were analyzed, all necessary measures were taken to ensure minimal sensitive data were exposed and that no de-anonymization was ever attempted. The results including crowd analyses were aimed at providing alternative ways to study human mobility while preserving individual privacy. In those studies, identifiers were only kept for a minimal amount of time at the hardware that was doing the sampling such that only counts and densities could ever be inferred, or until any other required metrics were computed. These careful efforts ensured not only that our work followed privacy and ethical guidelines, but could foster further similar research, such as privacy oriented services running at the Edge. We not only believe that more research like this is possible, but that moving forward, ethics takes a more central role in related studies and that commercial application only be implemented if certified to be following similar approaches.

To foster these discussions on privacy and ethics when using personal data from mobile users we include an extensive review to this thesis. In this review we observe that

in the past two decades, research on human mobility within networking has seen significant evolution, marked by a shift towards handling larger datasets (*i.e.*, more subjects, longer duration, and combined sources of data). This shift has been facilitated by automated data collection methods (*e.g.*, cellular network association, smartcards, phone tracking applications) and scalable analysis techniques (*e.g.*, big data and and advanced machine learning), which were previously unavailable. However, this rapid progress in human-centric research often raises concerns regarding privacy and ethical considerations. In this study [196], we conducted a comprehensive review of scientific literature on human mobility, with a particular focus on how ethical and privacy concerns have been addressed in their corresponding text. Our review encompassed 118 papers including a total of 149 individual mobility datasets. We found that while the expansion of data collections has paved the way for novel insights, adherence to established guidelines on data governance and transparent communication of research practices has not been consistent across studies. We conclude by initiating a discussion on the importance of addressing data governance, privacy, and ethical considerations within our research community, advocating for clearer guidelines and practices in these areas as well as having ethics a part of the formal education of future (computer) scientists.

In this thesis, we discussed results from two distinct but related perspectives in the interplay between human mobility and mobile devices. Firstly, we studied the relation between online activity (*e.g.*, browsing, Internet traffic, network connectivity, and gaming) and mobility (*e.g.*, visits, distances, and contacts). The quantification of these two aspects allowed us to model how online activity gives us a window into how humans move, change places, and come in contact with one another. As a consequence, through these models we were able to better predict movements and better able to predict the spread of information through a local community. This first group of results, when combined with large datasets and a responsible and ethical approach may guide the improvement of mobile apps and connectivity, as well as urban planning. Secondly, we studied how the omnipresence of online devices and connectivity allows us to study crowds, but may also compromised privacy by leaking mobility information. The study of large groups of people can help protect users privacy by no longer focusing (or even requiring) persistent identifiers, preventing any subject from being singled out. While wide spread Wi-Fi connectivity and always-ON devices may be convenient, we showed that changes in the Wi-Fi spectrum and Bluetooth device probing may be used to study large crowds in pre-determined spaces. From a privacy protection standpoint, our study of Bluetooth probing with location trackers from Apple revealed concerning potential leakage of location information of its users, such as location information and data exfiltration.

With that and the work presented in this thesis, we want to emphasize the following points among the remaining open research questions:

- *When subjects agree to it, what are less intrusive ways to capture mobility/behavioral data from mobile users?* Existing BLE trackers may provide a good source, even if not extremely accurate. Other less intrusive sources may include activity trackers (*e.g.*, watches and rings), driving telematics sensors, and satellite imagery.

- *How can mobile users ensure control over their data?* In order to deliver guarantees to users as well as to empower researchers when using sensitive data, reliable solutions to data sharing and data stewardship are needed and be widely adopted.
- *How can mobility research effectively communicate its findings while respecting and protecting the privacy of its subjects?* We believe both the research community and the involved subjects could benefit from guidelines into how to ethically (and legally) deal with private data. This way, we may keep the trust of the public and stand a chance at continuing to conduct research on human mobility with real data.

By addressing some of these open questions, another positive consequence is increased *reproducibility* of published results. This would be possible with sufficiently anonymized data being shared as well as new privacy-preserving methods that could be more easily reproduced. Therefore, to continue advancing our understanding of human behavior and its implications, new methods and sources of data are needed, where privacy and ethics discussions should be present from the outset.

Bibliography

[1] L. Tonetto, A. Carrara, A. Y. Ding, and J. Ott. Where is my tag? unveiling alternative uses of the apple findmy service. In *23rd IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks, WoWMoM 2022, Belfast, United Kingdom, June 14-17, 2022*, pages 396–405. IEEE, 2022. URL: https://doi.org/10.1109/WoWMoM54355.2022.00059, doi:10.1109/WoWMoM54355.2022.00059.

[2] B. Alipour, L. Tonetto, A. Y. Ding, R. Ketabi, J. Ott, and A. Helmy. Flutes vs. cellos: Analyzing mobility-traffic correlations in large WLAN traces. pages 1637–1645, 2018. URL: https://doi.org/10.1109/INFOCOM.2018.8486360, doi:10.1109/INFOCOM.2018.8486360.

[3] B. Alipour, L. Tonetto, R. Ketabi, A. Y. Ding, J. Ott, and A. Helmy. Where are you going next?: A practical multi-dimensional look at mobility prediction. In A. A. F. Loureiro, S. S. Kanhere, and P. Bellavista, editors, *Proceedings of the 22nd International ACM Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems, MSWiM 2019, Miami Beach, FL, USA, November 25-29, 2019*, pages 5–12. ACM, 2019. URL: https://doi.org/10.1145/3345768.3355923, doi:10.1145/3345768.3355923.

[4] L. Tonetto, M. Untersperger, and J. Ott. Towards exploiting wi-fi signals from low density infrastructure for crowd estimation. In S. Bayhan and E. Tsiropoulou, editors, *Proceedings of the 14th Workshop on Challenged Networks, CHANTS@MobiCom 2019, Los Cabos, Mexico, October 25, 2019*, pages 27–32. ACM, 2019. URL: https://doi.org/10.1145/3349625.3355439, doi:10.1145/3349625.3355439.

[5] L. Tonetto, E. Lagerspetz, A. Y. Ding, J. Ott, S. Tarkoma, and P. Nurmi. The mobility laws of location-based games. *EPJ Data Sci.*, 10(1):10, 2021. URL: https://doi.org/10.1140/epjds/s13688-021-00266-x, doi:10.1140/EPJDS/S13688-021-00266-X.

[6] L. Tonetto, M. Adikari, N. Mohan, A. Y. Ding, and J. Ott. Contact duration: Intricacies of human mobility. *Online Soc. Networks Media*, 28:100196, 2022. URL: https://doi.org/10.1016/j.osnem.2021.100196, doi:10.1016/j.osnem.2021.100196.

[7] P. Kister and L. Tonetto. On the importance of structural equivalence in temporal networks for epidemic forecasting. *Scientific Reports*, 13(1):866, 2023.

[8] V. D. Blondel, A. Decuyper, and G. Krings. A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4(1):1–55, 2015. URL: http://arxiv.org/abs/1502.03406, arXiv:1502.03406, doi:10.1140/epjds/s13688-015-0046-0.

[9] A. J. Oliner, A. P. Iyer, I. Stoica, E. Lagerspetz, and S. Tarkoma. Carat: collaborative energy diagnosis for mobile devices. In C. Petrioli, L. P. Cox, and K. Whitehouse, editors, *The 11th ACM Conference on Embedded Network Sensor Systems, SenSys '13, Roma, Italy, November 11-15, 2013*, pages 10:1–10:14. ACM, 2013. URL: https://doi.org/10.1145/2517351.2517354, doi:10.1145/2517351.2517354.

[10] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of human mobility on opportunistic forwarding algorithms. *IEEE Trans. Mob. Comput.*, 6(6):606–620, 2007. URL: https://doi.org/10.1109/TMC.2007.1060, doi:10.1109/TMC.2007.1060.

[11] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *nature*, 453(7196):779–782, 2008. doi:10.1038/nature06958.

[12] M. M. Ahamed and S. Faruque. 5g network coverage planning and analysis of the deployment challenges. *Sensors*, 21(19):6608, 2021.

[13] L. Pappalardo, L. Ferres, M. Sacasa, C. Cattuto, and L. Bravo. Evaluation of home detection algorithms on mobile phone data using individual-level ground truth. *EPJ Data Sci.*, 10(1):29, 2021. URL: https://doi.org/10.1140/epjds/s13688-021-00284-9, doi:10.1140/epjds/s13688-021-00284-9.

[14] R. Di Clemente, M. Luengo-Oroz, M. Travizano, S. Xu, B. Vaitla, and M. C. González. Sequences of purchases in credit card data reveal lifestyles in urban populations. *Nature communications*, 9(1):1–8, 2018.

[15] A. Draghici and M. van Steen. A survey of techniques for automatically sensing the behavior of a crowd. *ACM Comput. Surv.*, 51(1):21:1–21:40, 2018. URL: https://doi.org/10.1145/3129343, doi:10.1145/3129343.

[16] E. Fenske, D. Brown, J. Martin, T. Mayberry, P. Ryan, and E. C. Rye. Three years later: A study of MAC address randomization in mobile devices and when it succeeds. *Proc. Priv. Enhancing Technol.*, 2021(3):164–181, 2021. URL: https://doi.org/10.2478/popets-2021-0042, doi:10.2478/popets-2021-0042.

[17] N. Eagle and A. Pentland. Reality mining: sensing complex social systems. *Pers. Ubiquitous Comput.*, 10(4):255–268, 2006. URL: https://doi.org/10.1007/s00779-005-0046-3, doi:10.1007/s00779-005-0046-3.

[18] A. Stopczynski, V. Sekara, P. Sapiezynski, A. Cuttone, M. M. Madsen, J. E. Larsen, and S. Lehmann. Measuring large-scale social networks with high reso-

lution. *PLoS ONE*, 9(4), 2014. arXiv:1401.7233, doi:10.1371/journal.pone.0095978.

[19] R. Jurdak, K. Zhao, J. Liu, M. AbouJaoude, M. Cameron, and D. Newth. Understanding human mobility from Twitter. *PLoS ONE*, 10(7):35, 2015. URL: http://arxiv.org/abs/1412.2154https://www.scopus.com/inward/record.uri?eid=2-s2.0-84941367989&partnerID=40&md5=f3396e2e2175f3769c00097be58d5bb2, arXiv:arXiv:1412.2154v2, doi:10.1371/journal.pone.0131469.

[20] R. West, R. W. White, and E. Horvitz. Here and there: goals, activities, and predictions about location from geotagged queries. In G. J. F. Jones, P. Sheridan, D. Kelly, M. de Rijke, and T. Sakai, editors, *The 36th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '13, Dublin, Ireland - July 28 - August 01, 2013*, pages 817–820. ACM, 2013. URL: https://doi.org/10.1145/2484028.2484125, doi:10.1145/2484028.2484125.

[21] P. Callejo, M. Gramaglia, R. Cuevas, and Á. Cuevas. A deep dive into the accuracy of ip geolocation databases and its impact on online advertising. *IEEE Transactions on Mobile Computing*, pages 1–1, 2022. doi:10.1109/TMC.2022.3166785.

[22] M. Gouel, K. Vermeulen, R. Beverly, O. Fourmaux, and T. Friedman. IP geolocation database stability and implications for network research: A reproducibility study. In V. Bajpai, H. Haddadi, and O. Hohlfeld, editors, *5th Network Traffic Measurement and Analysis Conference, TMA 2021, Virtual Event, September 14-15, 2021*. IFIP, 2021. URL: http://dl.ifip.org/db/conf/tma/tma2021/tma2021-paper2.pdf.

[23] Apple. About privacy and location services in ios and ipados, 2022. URL: https://support.apple.com/en-gb/HT203033.

[24] Google. Manage your android device's location settings, 2022. URL: https://support.google.com/accounts/answer/3467281?hl=en.

[25] L. Sun, K. W. Axhausen, D.-H. Lee, and X. Huang. Understanding metropolitan patterns of daily encounters. *Proceedings of the National Academy of Sciences*, 110(34):13774–13779, 2013. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1306440110, arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1306440110, doi:10.1073/pnas.1306440110.

[26] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi. Limits of Predictability in Human Mobility. *Science*, 327(5968):1018–1021, 2010. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1177170, arXiv:0307014, doi:10.1126/science.1177170.

[27] S. Lu, J. Zhao, and H. Wang. Academic failures and co-location social networks in campus. *EPJ Data Science*, 11(1), 2022. URL: http://dx.doi.org/10.1140/epjds/s13688-022-00322-0, doi:10.1140/epjds/s13688-022-00322-0.

[28] G. Roussos, M. Musolesi, and G. D. Magoulas. Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. *Pervasive and Mobile Computing*, 6(4):455–466, 2010. Human Behavior in Ubiquitous Environments: Modeling of Human Mobility Patterns. URL: https://www.sciencedirect.com/science/article/pii/S1574119210000568, doi:https://doi.org/10.1016/j.pmcj.2010.07.002.

[29] J. Tang, F. Liu, Y. Wang, and H. Wang. Uncovering urban human mobility from large scale taxi gps data. *Physica A: Statistical Mechanics and its Applications*, 438:140–153, 2015. URL: https://www.sciencedirect.com/science/article/pii/S0378437115005853, doi:https://doi.org/10.1016/j.physa.2015.06.032.

[30] A. Rojas, P. Branch, and G. Armitage. Validation of the random waypoint mobility model through a real world mobility trace. pages 1–6, 2005.

[31] L. Pappalardo, G. Barlacchi, R. Pellungrini, and F. Simini. Human mobility from theory to practice: Data, models and applications. In S. Amer-Yahia, M. Mahdian, A. Goel, G. Houben, K. Lerman, J. J. McAuley, R. Baeza-Yates, and L. Zia, editors, *Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 1311–1312. ACM, 2019. URL: https://doi.org/10.1145/3308560.3320099, doi:10.1145/3308560.3320099.

[32] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, and M. Tomasini. Human mobility: Models and applications. *Physics Reports*, 734:1–74, 2018. Human mobility: Models and applications. URL: https://www.sciencedirect.com/science/article/pii/S037015731830022X, doi:https://doi.org/10.1016/j.physrep.2018.01.001.

[33] E. W. Montroll and G. H. Weiss. Random walks on lattices. ii. *Journal of Mathematical Physics*, 6(2):167–181, 1965.

[34] E. M. R. Oliveira, A. C. Viana, C. Sarraute, J. Brea, and J. I. Alvarez-Hamelin. On the regularity of human mobility. *Pervasive Mob. Comput.*, 33:73–90, 2016. URL: https://doi.org/10.1016/j.pmcj.2016.04.005, doi:10.1016/j.pmcj.2016.04.005.

[35] I. Rhee, M. Shin, S. Hong, K. Lee, S. J. Kim, and S. Chong. On the levy-walk nature of human mobility. *IEEE/ACM Trans. Netw.*, 19(3):630–643, 2011. URL: https://doi.org/10.1109/TNET.2011.2120618, doi:10.1109/TNET.2011.2120618.

[36] C. Song, T. Koren, P. Wang, and A.-L. Barabási. Modelling the scaling properties of human mobility. *Nature physics*, 6(10):818–823, 2010.

[37] L. Alessandretti, P. Sapiezynski, V. Sekara, S. Lehmann, and A. Baronchelli. Evidence for a conserved quantity in human mobility. *Nature human behaviour*, 2(7):485–491, 2018.

[38] F. Ekman, A. Keränen, J. Karvo, and J. Ott. Working day movement model. In M. Kim, C. Mascolo, and M. Musolesi, editors, *Proceedings of the 1st ACM SIGMOBILE workshop on Mobility Models, MobilityModels 2008, Hong Kong, China, May 26, 2008*, pages 33–40. ACM, 2008. URL: `https://doi.org/10.1145/1374688.1374695`, `doi:10.1145/1374688.1374695`.

[39] Q. Wang, N. E. Phillips, M. L. Small, and R. J. Sampson. Urban mobility and neighborhood isolation in america’s 50 largest cities. *Proceedings of the National Academy of Sciences*, 115(30):7735–7740, 2018. URL: `https://www.pnas.org/doi/abs/10.1073/pnas.1802537115`, `arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1802537115`, `doi:10.1073/pnas.1802537115`.

[40] S. Chang, E. Pierson, P. W. Koh, J. Gerardin, B. Redbird, D. Grusky, and J. Leskovec. Mobility network models of covid-19 explain inequities and inform reopening. *Nature*, 589(7840):82–87, 2021.

[41] L. Gauvin, M. Tizzoni, S. Piaggesi, A. Young, N. Adler, S. Verhulst, L. Ferres, and C. Cattuto. Gender gaps in urban mobility. *Humanities and Social Sciences Communications*, 7(1):1–13, 2020.

[42] A. Salgado, W. Li, F. Alhasoun, I. Caridi, and M. Gonzalez. Street context of various demographic groups in their daily mobility. *Applied Network Science*, 6(1):1–14, 2021.

[43] D. M. Powers. Applications and explanations of zipf's law. In *New methods in language processing and computational natural language learning*, 1998.

[44] K. Zhao et al. Explaining the power-law distribution of human mobility through transportation modality decomposition. *Scientific reports*, 5(1), 2015.

[45] L. Alessandretti, P. Sapiezynski, S. Lehmann, and A. Baronchelli. Multi-scale spatio-temporal analysis of human mobility. *PLOS ONE*, 12(2):1–17, 02 2017. URL: `https://doi.org/10.1371/journal.pone.0171686`, `doi:10.1371/journal.pone.0171686`.

[46] W. Wang et al. A comparative analysis of intra-city human mobility by taxi. *Physica A: Statistical Mechanics and its Applications*, 420, 2015.

[47] P. Sobkowicz et al. Lognormal distributions of user post lengths in internet discussions-a consequence of the weber-fechner law? *EPJ Data Science*, 2(1), 2013.

[48] C. Gros et al. Neuropsychological constraints to human data production on a global scale. *The European Physical Journal B*, 85, 2012.

[49] R. Gallotti et al. A stochastic model of randomly accelerated walkers for human mobility. *Nature communications*, 7(1), 2016.

[50] L. Gyarmati et al. Measuring user behavior in online social networks. *IEEE Network*, 24(5), 2010.

[51] N. Eikmeier et al. Revisiting power-law distributions in spectra of real world networks. In *Proceedings of ACM SIGKDD*, 2017.

[52] M. E. Newman. Clustering and preferential attachment in growing networks. *Physical review E*, 64(2), 2001.

[53] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In C. Apté, J. Ghosh, and P. Smyth, editors, *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011*, pages 1082–1090. ACM, 2011. URL: `https://doi.org/10.1145/2020408.2020579`, `doi:10.1145/2020408.2020579`.

[54] L. A. Adamic et al. Power-law distribution of the world wide web. *Science*, 287(5461), 2000.

[55] A. Clauset et al. Power-law distributions in empirical data. *SIAM review*, 51(4), 2009.

[56] R. Simard and P. L'Ecuyer. Computing the two-sided kolmogorov-smirnov distribution. *Journal of Statistical Software*, 39:1–18, 2011.

[57] C. E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423, 1948. URL: `https://doi.org/10.1002/j.1538-7305.1948.tb01338.x`, `doi:10.1002/j.1538-7305.1948.tb01338.x`.

[58] Y. Gao, I. Kontoyiannis, and E. Bienenstock. Estimating the entropy of binary time series: Methodology, some theory and a simulation study. *Entropy*, 10(2):71–99, 2008.

[59] T. Cover and J. Thomas. Elements of Information Theory., 1991.

[60] A. J. Wyner and D. Foster. On the lower limits of entropy estimation. *IEEE Transactions on Information Theory*, pages 1–19, 2003.

[61] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Trans. Inf. Theory*, 23(3):337–343, 1977. URL: `https://doi.org/10.1109/TIT.1977.1055714`, `doi:10.1109/TIT.1977.1055714`.

[62] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens. The context-tree weighting method: basic properties. *IEEE Trans. Inf. Theory*, 41(3):653–664, 1995. URL: `https://doi.org/10.1109/18.382012`, `doi:10.1109/18.382012`.

[63] M. Burrows and D. J. Wheeler. A block-sorting lossless data compression algorithm. *Digital SRC*, 1994.

[64] H. Cai, S. R. Kulkarni, and S. Verdú. Universal entropy estimation via block sorting. *IEEE Trans. Inf. Theory*, 50(7):1551–1561, 2004. URL: `https://doi.org/10.1109/TIT.2004.830771`, `doi:10.1109/TIT.2004.830771`.

[65] E. M. McCreight. A space-economical suffix tree construction algorithm. *J. ACM*, 23(2):262–272, 1976. URL: `https://doi.org/10.1145/321941.321946`, `doi:10.1145/321941.321946`.

[66] D. Adjeroh, T. Bell, and A. Mukherjee. *The Burrows-Wheeler Transform: Data Compression, Suffix Arrays, and Pattern Matching*. Springer Science & Business Media, 2008.

[67] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.

[68] X. Liang, J. Zhao, L. Dong, and K. Xu. Unraveling the origin of exponential law in intra-urban human mobility. *Scientific reports*, 3(1):1–7, 2013.

[69] X. Lu, L. Bengtsson, and P. Holme. Predictability of population displacement after the 2010 haiti earthquake. *Proceedings of the National Academy of Sciences*, 109(29):11576–11581, 2012.

[70] M. U. Kraemer, A. Sadilek, Q. Zhang, N. A. Marchal, G. Tuli, E. L. Cohn, Y. Hswen, T. A. Perkins, D. L. Smith, R. C. Reiner, et al. Mapping global variation in human mobility. *Nature Human Behaviour*, 4(8):800–810, 2020.

[71] P. Hui, J. Crowcroft, and E. Yoneki. Bubble rap: Social-based forwarding in delay-tolerant networks. *IEEE transactions on mobile computing*, 10(11):1576–1589, 2010.

[72] S. Flaxman, S. Mishra, A. Gandy, H. J. T. Unwin, T. A. Mellan, H. Coupland, C. Whittaker, H. Zhu, T. Berah, J. W. Eaton, et al. Estimating the effects of non-pharmaceutical interventions on covid-19 in europe. *Nature*, 584(7820):257–261, 2020.

[73] M. U. Kraemer, C.-H. Yang, B. Gutierrez, C.-H. Wu, B. Klein, D. M. Pigott, O. C.-. D. W. Group†, L. Du Plessis, N. R. Faria, R. Li, et al. The effect of human mobility and control measures on the covid-19 epidemic in china. *Science*, 368(6490):493–497, 2020.

[74] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and C. Diot. Pocket switched networks and human mobility in conference environments. In K. Fall and S. Keshav, editors, *Proceedings of the 2005 ACM SIGCOMM Workshop on Delay-Tolerant Networking, WDTN '05, Philadelphia, Pennsylvania, USA, August 26, 2005*, pages 244–251. ACM, 2005. URL: `https://doi.org/10.1145/1080139.1080142`, `doi:10.1145/1080139.1080142`.

[75] G. Maier, F. Schneider, and A. Feldmann. A first look at mobile hand-held device traffic. In A. Krishnamurthy and B. Plattner, editors, *Passive and Active Measurement, 11th International Conference, PAM 2010, Zurich, Switzerland, April 7-9, 2010. Proceedings*, volume 6032 of *Lecture Notes in Computer Science*, pages 161–170. Springer, 2010. URL: https://doi.org/10.1007/978-3-642-12334-4_17, doi:10.1007/978-3-642-12334-4_17.

[76] U. Kumar, J. Kim, and A. Helmy. Changing patterns of mobile network (WLAN) usage: Smart-phones vs. laptops. In R. Saracco, K. B. Letaief, M. Gerla, S. Palazzo, and L. Atzori, editors, *2013 9th International Wireless Communications and Mobile Computing Conference, IWCMC 2013, Sardinia, Italy, July 1-5, 2013*, pages 1584–1589. IEEE, 2013. URL: https://doi.org/10.1109/IWCMC.2013.6583792, doi:10.1109/IWCMC.2013.6583792.

[77] X. Chen, R. Jin, K. Suh, B. Wang, and W. Wei. Network performance of smart mobile handhelds in a university campus wifi network. In J. W. Byers, J. Kurose, R. Mahajan, and A. C. Snoeren, editors, *Proceedings of the 12th ACM SIGCOMM Internet Measurement Conference, IMC '12, Boston, MA, USA, November 14-16, 2012*, pages 315–328. ACM, 2012. URL: https://doi.org/10.1145/2398776.2398809, doi:10.1145/2398776.2398809.

[78] A. Gember, A. Anand, and A. Akella. A comparative study of handheld and non-handheld traffic in campus wi-fi networks. In N. Spring and G. F. Riley, editors, *Passive and Active Measurement - 12th International Conference, PAM 2011, Atlanta, GA, USA, March 20-22, 2011. Proceedings*, volume 6579 of *Lecture Notes in Computer Science*, pages 173–183. Springer, 2011. URL: https://doi.org/10.1007/978-3-642-19260-9_18, doi:10.1007/978-3-642-19260-9_18.

[79] M. Afanasyev, T. Chen, G. M. Voelker, and A. C. Snoeren. Analysis of a mixed-use urban wifi network: when metropolitan becomes neapolitan. In K. Papagiannaki and Z. Zhang, editors, *Proceedings of the 8th ACM SIGCOMM Internet Measurement Conference, IMC 2008, Vouliagmeni, Greece, October 20-22, 2008*, pages 85–98. ACM, 2008. URL: https://doi.org/10.1145/1452520.1452531, doi:10.1145/1452520.1452531.

[80] I. Papapanagiotou, E. M. Nahum, and V. Pappas. Smartphones vs. laptops: comparing web browsing behavior and the implications for caching. In P. G. Harrison, M. F. Arlitt, and G. Casale, editors, *ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems, SIGMETRICS '12, London, United Kingdom, June 11-15, 2012*, pages 423–424. ACM, 2012. URL: https://doi.org/10.1145/2254756.2254824, doi:10.1145/2254756.2254824.

[81] H. Falaki, D. Lymberopoulos, R. Mahajan, S. Kandula, and D. Estrin. A first look at traffic on smartphones. In M. Allman, editor, *Proceedings of the 10th ACM SIGCOMM Internet Measurement Conference, IMC 2010, Melbourne, Australia*

- *November 1-3, 2010*, pages 281–287. ACM, 2010. URL: https://doi.org/10.1145/1879141.1879176, doi:10.1145/1879141.1879176.

[82] A. K. Das, P. H. Pathak, C. Chuah, and P. Mohapatra. Characterization of wireless multidevice users. *ACM Trans. Internet Techn.*, 16(4):29:1–29:25, 2016. URL: https://doi.org/10.1145/2955096, doi:10.1145/2955096.

[83] P. Y. Cao, G. Li, A. C. Champion, D. Xuan, S. Romig, and W. Zhao. On human mobility predictability via WLAN logs. In *2017 IEEE Conference on Computer Communications, INFOCOM 2017, Atlanta, GA, USA, May 1-4, 2017*, pages 1–9. IEEE, 2017. URL: https://doi.org/10.1109/INFOCOM.2017.8057234, doi:10.1109/INFOCOM.2017.8057234.

[84] S. Moghaddam and A. Helmy. Multidimensional modeling and analysis of wireless users online activity and mobility: a neural-networks map approach. In A. Helmy, B. Landfeldt, and L. Bononi, editors, *Proceedings of the 14th International Symposium on Modeling Analysis and Simulation of Wireless and Mobile Systems, MSWiM 2011, Miami, Florida, USA, October 31 - November 4, 2011*, pages 401–408. ACM, 2011. URL: https://doi.org/10.1145/2068897.2068965, doi:10.1145/2068897.2068965.

[85] T. Althoff, R. W. White, E. Horvitz, et al. Influence of pokémon go on physical activity: study and implications. *Journal of medical Internet research*, 18(12):e6759, 2016.

[86] H. Xu, Y. Xian, H. Xu, L. Liang, A. F. Hernandez, T. Y. Wang, and E. D. Peterson. Does pokémon go help players be more active? an evaluation of pokémon go and physical activity. *Circulation*, 135(suppl_1):A02–A02, 2017.

[87] M. Khamzina, K. V. Parab, R. An, T. Bullard, and D. S. Grigsby-Toussaint. Impact of pokémon go on physical activity: A systematic review and meta-analysis. *American Journal of Preventive Medicine*, 58(2):270–282, 2020.

[88] S. A. Lear, W. Hu, S. Rangarajan, D. Gasevic, D. Leong, R. Iqbal, A. Casanova, S. Swaminathan, R. M. Anjana, R. Kumar, et al. The effect of physical activity on mortality and cardiovascular disease in 130 000 people from 17 high-income, middle-income, and low-income countries: the pure study. *The Lancet*, 390(10113):2643–2654, 2017.

[89] K. Hu, R. F. Riemersma-Van Der Lek, M. Patxot, P. Li, S. A. Shea, F. A. Scheer, and E. J. Van Someren. Progression of dementia assessed by temporal correlations of physical activity: results from a 3.5-year, longitudinal randomized controlled trial. *Scientific Reports*, 6(1):1–10, 2016.

[90] T. Althoff, P. Jindal, and J. Leskovec. Online actions with offline impact: How online social networks influence online and offline user behavior. In M. de Rijke, M. Shokouhi, A. Tomkins, and M. Zhang, editors, *Proceedings of the Tenth ACM*

International Conference on Web Search and Data Mining, WSDM 2017, Cambridge, United Kingdom, February 6-10, 2017, pages 537–546. ACM, 2017. URL: https://doi.org/10.1145/3018661.3018672, doi:10.1145/3018661.3018672.

[91] F. Y. Wong. Influence of Pokémon Go on physical activity levels of university players: a cross-sectional study. *International Journal of Health Geographics*, 16(1):17, 2017.

[92] R. Gal, A. M. May, E. J. van Overmeeren, M. Simons, and E. M. Monninkhof. The effect of physical activity interventions comprising wearables and smartphone applications on physical activity: a systematic review and meta-analysis. *Sports medicine-open*, 4(1):42, 2018.

[93] A. Shameli, T. Althoff, A. Saberi, and J. Leskovec. How gamification affects physical activity: Large-scale analysis of walking challenges in a mobile application. In R. Barrett, R. Cummings, E. Agichtein, and E. Gabrilovich, editors, *Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017*, pages 455–463. ACM, 2017. URL: https://doi.org/10.1145/3041021.3054172, doi:10.1145/3041021.3054172.

[94] A. Romeo, S. Edney, R. Plotnikoff, R. Curtis, J. Ryan, I. Sanders, A. Crozier, and C. Maher. Can smartphone apps increase physical activity? systematic review and meta-analysis. *Journal of medical Internet research*, 21(3):e12053, 2019.

[95] G. Smith, R. Wieser, J. Goulding, and D. Barrack. A refined limit on the predictability of human mobility. In *IEEE International Conference on Pervasive Computing and Communications, PerCom 2014, Budapest, Hungary, March 24-28, 2014*, pages 88–94. IEEE Computer Society, 2014. URL: https://doi.org/10.1109/PerCom.2014.6813948, doi:10.1109/PerCom.2014.6813948.

[96] Y. Li, D. Jin, P. Hui, Z. Wang, and S. Chen. Limits of predictability for large-scale urban vehicular mobility. *IEEE Trans. Intell. Transp. Syst.*, 15(6):2671–2682, 2014. URL: https://doi.org/10.1109/TITS.2014.2325395, doi:10.1109/TITS.2014.2325395.

[97] J. Wang, Y. Mao, J. Li, Z. Xiong, and W. X. Wang. Predictability of road traffic and congestion in urban areas. *PLoS ONE*, 2015.

[98] R. Gallotti, A. Bazzani, M. D. Esposti, and S. Rambaldi. Entropic measures of individual mobility patterns. *JSTAT*, 2013.

[99] T. Takaguchi, M. Nakamura, N. Sato, K. Yano, and N. Masuda. Predictability of conversation partners. *Physical Review X*, 2011.

[100] R. Sinatra and M. Szell. Entropy and the predictability of online life. *Entropy*, 16(1):543–556, 2014.

[101] R. Hanel and S. Thurner. A comprehensive classification of complex statistical systems and an axiomatic derivation of their entropy and distribution functions. *Epl*, 93(2), 2011.

[102] X. Zhou, Z. Zhao, R. Li, Y. Zhou, and H. Zhang. The predictability of cellular networks traffic. In *International Symposium on Communications and Information Technologies, ISCIT 2012, Gold Coast, Australia, October 2-5, 2012*, pages 973–978. IEEE, 2012. URL: https://doi.org/10.1109/ISCIT.2012.6381046, doi:10.1109/ISCIT.2012.6381046.

[103] G. Goulet-Langlois, H. N. Koutsopoulos, Z. Zhao, and J. Zhao. Measuring regularity of individual travel patterns. *IEEE Trans. Intell. Transp. Syst.*, 19(5):1583–1592, 2018. URL: https://doi.org/10.1109/TITS.2017.2728704, doi:10.1109/TITS.2017.2728704.

[104] M. Salathé, M. Kazandjieva, J. W. Lee, P. Levis, M. W. Feldman, and J. H. Jones. A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences*, 107(51):22020–22025, 2010. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1009094108, arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1009094108, doi:10.1073/pnas.1009094108.

[105] L. Isella et al. What's in a crowd? analysis of face-to-face behavioral networks. *Journal of theoretical biology*, 271(1):166–180, 2011.

[106] T. Hossmann, T. Spyropoulos, and F. Legendre. Putting contacts into context: mobility modeling beyond inter-contact times. In *Proceedings of the 12th ACM International Symposium on Mobile Ad Hoc Networking and Computing, MobiHoc 2011, Paris, France, May 16-20, 2011*, page 18. ACM, 2011. URL: https://doi.org/10.1145/2107502.2107526, doi:10.1145/2107502.2107526.

[107] K. A. Fisher, M. W. Tenforde, L. R. Feldstein, C. J. Lindsell, N. I. Shapiro, D. C. Files, K. W. Gibbs, H. L. Erickson, M. E. Prekker, J. S. Steingrub, et al. Community and close contact exposures associated with covid-19 among symptomatic adults ¿= 18 years in 11 outpatient health care facilities—united states, july 2020. *Morbidity and Mortality Weekly Report*, 69(36):1258, 2020.

[108] N. Masuda and P. Holme. Predicting and controlling infectious disease epidemics using temporal networks. *F1000prime reports*, 5, 2013.

[109] K. Sato, M. Oka, A. Barrat, and C. Cattuto. Predicting partially observed processes on temporal networks by dynamics-aware node embeddings (DyANE). *EPJ Data Science*, 10(1), may 2021. URL: https://doi.org/10.1140%2Fepjds%2Fs13688-021-00277-8, doi:10.1140/epjds/s13688-021-00277-8.

[110] B. Zhang, P. A. Pavlou, and R. Krishnan. On direct vs. indirect peer influence in large social networks. *Information Systems Research*, 29(2):292–314,

2018. URL: https://doi.org/10.1287/isre.2017.0753, arXiv:https://doi.org/10.1287/isre.2017.0753, doi:10.1287/isre.2017.0753.

[111] F. Musciotto and S. Miccichè. Effective strategies for targeted attacks to the network of cosa nostra affiliates. *EPJ Data Science*, 11(1):11, Feb 2022. URL: https://doi.org/10.1140/epjds/s13688-022-00323-z, doi:10.1140/epjds/s13688-022-00323-z.

[112] R. A. Rossi and N. K. Ahmed. Role discovery in networks. *CoRR*, abs/1405.7134, 2014. URL: http://arxiv.org/abs/1405.7134, arXiv:1405.7134.

[113] L. Wang, C. Huang, W. Ma, Y. Lu, and S. Vosoughi. Embedding node structural role identity using stress majorization. *CoRR*, abs/2109.07023, 2021. URL: https://arxiv.org/abs/2109.07023, arXiv:2109.07023.

[114] D. R. Figueiredo, L. F. R. Ribeiro, and P. H. P. Saverese. struc2vec: Learning node representations from structural identity. *CoRR*, abs/1704.03165, 2017. URL: http://arxiv.org/abs/1704.03165, arXiv:1704.03165.

[115] K. Tu, P. Cui, X. Wang, P. S. Yu, and W. Zhu. Deep recursive network embedding with regular equivalence. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 2357–2366, New York, NY, USA, 2018. Association for Computing Machinery. URL: https://doi.org/10.1145/3219819.3220068, doi:10.1145/3219819.3220068.

[116] J. Jin, M. Heimann, D. Jin, and D. Koutra. Toward understanding and evaluating structural node embeddings. *ACM Trans. Knowl. Discov. Data*, 16(3), nov 2021. URL: https://doi.org/10.1145/3481639, doi:10.1145/3481639.

[117] F. Schliski, J. Schlötterer, and M. Granitzer. Influence of random walk parametrization on graph embeddings. In J. M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M. J. Silva, and F. Martins, editors, *Advances in Information Retrieval*, pages 58–65, Cham, 2020. Springer International Publishing.

[118] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[119] M. A. Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[120] E. Peltonen, E. Lagerspetz, P. Nurmi, and S. Tarkoma. Energy modeling of system settings: A crowdsourced approach. In *2015 IEEE International Conference on*

Pervasive Computing and Communications, PerCom 2015, St. Louis, MO, USA, 23-27 March, 2015, pages 37–45. IEEE Computer Society, 2015. URL: https://doi.org/10.1109/PERCOM.2015.7146507, doi:10.1109/PERCOM.2015.7146507.

[121] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*, pages 280–289. AAAI Press, 2017. URL: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15587.

[122] P. Rasche, A. Schlomann, and A. Mertens. Who Is Still Playing Pokémon Go? A Web-Based Survey. *JMIR Serious Games*, 5(2):e7, 2017.

[123] A. Colley, J. Thebault-Spieker, A. Y. Lin, D. Degraen, B. Fischman, J. Häkkilä, K. Kuehl, V. Nisi, N. J. Nunes, N. Wenig, D. Wenig, B. J. Hecht, and J. Schöning. The geography of pokémon GO: beneficial and problematic effects on places and movement. In G. Mark, S. R. Fussell, C. Lampe, m. c. schraefel, J. P. Hourcade, C. Appert, and D. Wigdor, editors, *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, May 06-11, 2017*, pages 1179–1192. ACM, 2017. URL: https://doi.org/10.1145/3025453.3025495, doi:10.1145/3025453.3025495.

[124] M. Ester, H. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In E. Simoudis, J. Han, and U. M. Fayyad, editors, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA*, pages 226–231. AAAI Press, 1996. URL: http://www.aaai.org/Library/KDD/1996/kdd96-037.php.

[125] D. Ashbrook and T. Starner. Using GPS to learn significant locations and predict movement across multiple users. *Pers. Ubiquitous Comput.*, 7(5):275–286, 2003. URL: https://doi.org/10.1007/s00779-003-0240-0, doi:10.1007/s00779-003-0240-0.

[126] G. Goulet-Langlois, H. N. Koutsopoulos, Z. Zhao, and J. Zhao. Measuring regularity of individual travel patterns. *IEEE Trans. Intell. Transp. Syst.*, 19(5):1583–1592, 2018. URL: https://doi.org/10.1109/TITS.2017.2728704, doi:10.1109/TITS.2017.2728704.

[127] D. S. Candido et al. Evolution and epidemic spread of sars-cov-2 in brazil. *Science*, 369(6508), 2020.

[128] S. M. Kissler et al. Reductions in commuting mobility correlate with geographic differences in sars-cov-2 prevalence in new york city. *Nature communications*, 11(1), 2020.

[129] S. Chang et al. Mobility network models of covid-19 explain inequities and inform reopening. *Nature*, 2020.

[130] K. Soltesz et al. The effect of interventions on covid-19. *Nature*, 588(7839), 2020.

[131] C. Cattuto et al. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PloS one*, 5(7), 2010.

[132] L. Ferretti et al. Quantifying sars-cov-2 transmission suggests epidemic control with digital contact tracing. *Science*, 368(6491), 2020.

[133] D. Ferreira, V. Kostakos, and A. K. Dey. AWARE: mobile context instrumentation framework. *Frontiers ICT*, 2:6, 2015. URL: `https://doi.org/10.3389/fict.2015.00006`, `doi:10.3389/fict.2015.00006`.

[134] M. Musolesi and C. Mascolo. A community based mobility model for ad hoc network research. In M. Conti, J. Crowcroft, and A. Passarella, editors, *Proceedings of the 2nd International Workshop on Multi-Hop Ad Hoc Networks: From Theory to Reality, REALMAN@MobiHoc 2006, Florence, Italy, May 26, 2006*, pages 31–38. ACM, 2006. URL: `https://doi.org/10.1145/1132983.1132990`, `doi:10.1145/1132983.1132990`.

[135] Y. Zheng, X. Xie, and W. Ma. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33(2):32–39, 2010. URL: `http://sites.computer.org/debull/A10june/geolife.pdf`.

[136] Y. Zheng, L. Zhang, X. Xie, and W. Ma. Mining interesting locations and travel sequences from GPS trajectories. In J. Quemada, G. León, Y. S. Maarek, and W. Nejdl, editors, *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*, pages 791–800. ACM, 2009. URL: `https://doi.org/10.1145/1526709.1526816`, `doi:10.1145/1526709.1526816`.

[137] P. Widhalm, Y. Yang, M. Ulm, S. Athavale, and M. C. González. Discovering urban activity patterns in cell phone data. *Transportation*, 42(4):597–623, 2015.

[138] L. Alessandretti et al. Multi-scale spatio-temporal analysis of human mobility. *PloS one*, 12(2), 2017.

[139] N. Askitas, K. Tatsiramos, and B. Verheyden. Estimating worldwide effects of non-pharmaceutical interventions on covid-19 incidence and population mobility patterns using a multiple-event study. *Scientific reports*, 11(1):1–13, 2021.

[140] M. U. G. Kraemer, C.-H. Yang, B. Gutierrez, C.-H. Wu, B. Klein, D. M. Pigott, null null, L. du Plessis, N. R. Faria, R. Li, W. P. Hanage, J. S. Brownstein, M. Layan, A. Vespignani, H. Tian, C. Dye, O. G. Pybus, and S. V. Scarpino. The effect of human mobility and control measures on the covid-19 epidemic in china. *Science*, 368(6490):493–497, 2020. URL: `https://www.science.org/doi/abs/10.1126/science.abb4218`, `arXiv:https://www.science.org/doi/pdf/10.1126/science.abb4218`, `doi:10.1126/science.abb4218`.

[141] L. Tian, X. Li, F. Qi, Q.-Y. Tang, V. Tang, J. Liu, Z. Li, X. Cheng, X. Li, Y. Shi, H. Liu, and L.-H. Tang. Harnessing peak transmission around symptom onset for non-pharmaceutical intervention and containment of the covid-19 pandemic. *Nature Communications*, 12(1):1147, Feb 2021. URL: https://doi.org/10.1038/s41467-021-21385-z, doi:10.1038/s41467-021-21385-z.

[142] Y. Ge, W.-B. Zhang, X. Wu, C. W. Ruktanonchai, H. Liu, J. Wang, Y. Song, M. Liu, W. Yan, J. Yang, E. Cleary, S. H. Qader, F. Atuhaire, N. W. Ruktanonchai, A. J. Tatem, and S. Lai. Untangling the changing impact of non-pharmaceutical interventions and vaccination on european covid-19 trajectories. *Nature Communications*, 13(1):3106, Jun 2022. URL: https://doi.org/10.1038/s41467-022-30897-1, doi:10.1038/s41467-022-30897-1.

[143] C. Xiong, S. Hu, M. Yang, W. Luo, and L. Zhang. Mobile device data reveal the dynamics in a positive relationship between human mobility and covid-19 infections. *Proceedings of the National Academy of Sciences*, 117(44):27087–27089, 2020. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2010836117, arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2010836117, doi:10.1073/pnas.2010836117.

[144] E. Pinto, E. Nepomuceno, and A. Campanharo. Impact of network topology on the spread of infectious diseases. *TEMA*, 21:95–115, 04 2020. doi:10.5540/tema.2020.021.01.0095.

[145] C. Fan, R. Lee, Y. Yang, and A. Mostafavi. Fine-grained data reveal segregated mobility networks and opportunities for local containment of covid-19. *Scientific Reports*, 11(1):16895, Aug 2021. URL: https://doi.org/10.1038/s41598-021-95894-8, doi:10.1038/s41598-021-95894-8.

[146] Y. Moreno and A. Vazquez. Disease spreading in structured scale-free networks. *Physics of Condensed Matter*, 31, 10 2002. doi:10.1140/epjb/e2003-00031-9.

[147] A. Grover and J. Leskovec. Node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 855–864, New York, NY, USA, 2016. Association for Computing Machinery. URL: https://doi.org/10.1145/2939672.2939754, doi:10.1145/2939672.2939754.

[148] M. S. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973. URL: https://doi.org/10.1086/225469, arXiv:https://doi.org/10.1086/225469, doi:10.1086/225469.

[149] N. Eagle and A. (Sandy) Pentland. Reality mining: Sensing complex social systems. *Personal Ubiquitous Comput.*, 10(4):255–268, 03 2006. URL: https://doi.org/10.1007/s00779-005-0046-3, doi:10.1007/s00779-005-0046-3.

[150] M. Génois and A. Barrat. Can co-location be used as a proxy for face-to-face contacts? *EPJ Data Science*, 7(1):11, 05 2018. URL: https://doi.org/10.1140/epjds/s13688-018-0140-1, doi:10.1140/epjds/s13688-018-0140-1.

[151] D. Rybski, S. V. Buldyrev, S. Havlin, F. Liljeros, and H. A. Makse. Scaling laws of human interaction activity. *Proceedings of the National Academy of Sciences*, 106(31):12640–12645, 2009. URL: https://www.pnas.org/doi/abs/10.1073/pnas.0902667106, arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.0902667106, doi:10.1073/pnas.0902667106.

[152] S. Gounane, Y. Barkouch, A. Atlas, M. Bendahmane, F. Karami, and D. Meskine. An adaptive social distancing sir model for covid-19 disease spreading and forecasting. *Epidemiologic Methods*, 10(s1):20200044, 2021. URL: https://doi.org/10.1515/em-2020-0044 [last checked 2022-06-21], doi:doi:10.1515/em-2020-0044.

[153] K. Tantrakarnapa, B. Bhopdhornangkul, and K. Nakhaapakorn. Influencing factors of covid-19 spreading: a case study of thailand. *Journal of Public Health*, 30(3):621–627, Mar 2022. URL: https://doi.org/10.1007/s10389-020-01329-5, doi:10.1007/s10389-020-01329-5.

[154] X. Qian, L. Sun, and S. V. Ukkusuri. Scaling of contact networks for epidemic spreading in urban transit systems. *Scientific Reports*, 11(1):4408, Feb 2021. URL: https://doi.org/10.1038/s41598-021-83878-7, doi:10.1038/s41598-021-83878-7.

[155] M. H. Riad, M. Sekamatte, F. Ocom, I. Makumbi, and C. M. Scoglio. Risk assessment of ebola virus disease spreading in uganda using a two-layer temporal network. *Scientific Reports*, 9(1):16060, Nov 2019. URL: https://doi.org/10.1038/s41598-019-52501-1, doi:10.1038/s41598-019-52501-1.

[156] Y. A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports*, 3:1–5, 2013. doi:10.1038/srep01376.

[157] E. Commission. Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, 2021.

[158] S. of California Department of Justice. California consumer privacy act (CCPA), 2018. URL: https://oag.ca.gov/privacy/ccpa.

[159] B. Bonné, A. Barzan, P. Quax, and W. Lamotte. Wifipi: Involuntary tracking of visitors at mass events. In *IEEE 14th International Symposium on "A World of Wireless, Mobile and Multimedia Networks", WoWMoM 2013, Madrid, Spain, June 4-7, 2013*, pages 1–6. IEEE Computer Society, 2013. URL: https://doi.org/10.1109/WoWMoM.2013.6583443, doi:10.1109/WoWMoM.2013.6583443.

[160] Y. Fukuzaki, N. Nishio, M. Mochizuki, and K. Murao. A pedestrian flow analysis system using wi-fi packet sensors to a real environment. In A. J. Brush, A. Friday, J. A. Kientz, J. Scott, and J. Song, editors, *The 2014 ACM Conference on Ubiquitous Computing, UbiComp '14 Adjunct, Seattle, WA, USA - September 13 - 17, 2014*, pages 721–730. ACM, 2014. URL: https://doi.org/10.1145/2638728.2641312, doi:10.1145/2638728.2641312.

[161] L. Schauer, M. Werner, and P. Marcus. Estimating crowd densities and pedestrian flows using wi-fi and bluetooth. In M. Youssef, C. Mascolo, and F. Kawsar, editors, *11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, MOBIQUITOUS 2014, London, United Kingdom, December 2-5, 2014*, pages 171–177. ICST / ACM, 2014. URL: https://doi.org/10.4108/icst.mobiquitous.2014.257870, doi:10.4108/icst.mobiquitous.2014.257870.

[162] A. B. M. Musa and J. Eriksson. Tracking unmodified smartphones using wi-fi monitors. In M. R. Eskicioglu, A. Campbell, and K. Langendoen, editors, *The 10th ACM Conference on Embedded Network Sensor Systems, SenSys '12, Toronto, ON, Canada, November 6-9, 2012*, pages 281–294. ACM, 2012. URL: https://doi.org/10.1145/2426656.2426685, doi:10.1145/2426656.2426685.

[163] E. Vattapparamban, B. S. Ciftler, I. Güvenç, K. Akkaya, and A. Kadri. Indoor occupancy tracking in smart buildings using passive sniffing of probe requests. In *IEEE International Conference on Communication, ICC 2015, London, United Kingdom, June 8-12, 2015, Workshop Proceedings*, pages 38–44. IEEE, 2016. URL: https://doi.org/10.1109/ICCW.2016.7503761, doi:10.1109/ICCW.2016.7503761.

[164] L. M. Mikkelsen, R. Buchakchiev, T. K. Madsen, and H. Schwefel. Public transport occupancy estimation using WLAN probing. In *2016 8th International Workshop on Resilient Networks Design and Modeling (RNDM), Halmstad, Sweden, September 13-15, 2016*, pages 302–308. IEEE, 2016. URL: https://doi.org/10.1109/RNDM.2016.7608302, doi:10.1109/RNDM.2016.7608302.

[165] M. V. Barbera, A. Epasto, A. Mei, V. C. Perta, and J. Stefa. Signals from the crowd: uncovering social relationships through smartphone probes. In K. Papagiannaki, P. K. Gummadi, and C. Partridge, editors, *Proceedings of the 2013 Internet Measurement Conference, IMC 2013, Barcelona, Spain, October 23-25, 2013*, pages 265–276. ACM, 2013. URL: https://doi.org/10.1145/2504730.2504742, doi:10.1145/2504730.2504742.

[166] H. Hong, C. Luo, and M. C. Chan. Socialprobe: Understanding social interaction through passive wifi monitoring. In T. Hara and H. Shigeno, editors, *Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, MobiQuitous 2016, Hiroshima, Japan, November 28 -*

December 1, 2016, pages 94–103. ACM, 2016. URL: https://doi.org/10.1145/2994374.2994387, doi:10.1145/2994374.2994387.

[167] K. Chintalapudi, A. P. Iyer, and V. N. Padmanabhan. Indoor localization without the pain. In N. H. Vaidya, S. Banerjee, and D. Katabi, editors, *Proceedings of the 16th Annual International Conference on Mobile Computing and Networking, MOBICOM 2010, Chicago, Illinois, USA, September 20-24, 2010*, pages 173–184. ACM, 2010. URL: https://doi.org/10.1145/1859995.1860016, doi:10.1145/1859995.1860016.

[168] M. Seifeldin, A. Saeed, A. E. Kosba, A. El-Keyi, and M. Youssef. Nuzzer: A large-scale device-free passive localization system for wireless environments. *IEEE Trans. Mob. Comput.*, 12(7):1321–1334, 2013. URL: https://doi.org/10.1109/TMC.2012.106, doi:10.1109/TMC.2012.106.

[169] S. Depatla and Y. Mostofi. Crowd counting through walls using wifi. In *2018 IEEE International Conference on Pervasive Computing and Communications, PerCom 2018, Athens, Greece, March 19-23, 2018*, pages 1–10. IEEE Computer Society, 2018. URL: https://doi.org/10.1109/PERCOM.2018.8444589, doi:10.1109/PERCOM.2018.8444589.

[170] W. Xi, J. Zhao, X. Li, K. Zhao, S. Tang, X. Liu, and Z. Jiang. Electronic frog eye: Counting crowd using wifi. In *2014 IEEE Conference on Computer Communications, INFOCOM 2014, Toronto, Canada, April 27 - May 2, 2014*, pages 361–369. IEEE, 2014. URL: https://doi.org/10.1109/INFOCOM.2014.6847958, doi:10.1109/INFOCOM.2014.6847958.

[171] S. Palipana, P. Agrawal, and D. Pesch. Channel state information based human presence detection using non-linear techniques. In *Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments, BuildSys@SenSys 2016, Palo Alto, CA, USA, November 16-17, 2016*, pages 177–186. ACM, 2016. URL: https://doi.org/10.1145/2993422.2993579, doi:10.1145/2993422.2993579.

[172] C. A. Pouw, F. Toschi, F. van Schadewijk, and A. Corbetta. Monitoring physical distancing for crowd management: Real-time trajectory and group analysis. *PloS one*, 15(10):e0240963, 2020.

[173] L. Boominathan, S. S. S. Kruthiventi, and R. V. Babu. Crowdnet: A deep convolutional network for dense crowd counting. In A. Hanjalic, C. Snoek, M. Worring, D. C. A. Bulterman, B. Huet, A. Kelliher, Y. Kompatsiaris, and J. Li, editors, *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016*, pages 640–644. ACM, 2016. URL: https://doi.org/10.1145/2964284.2967300, doi:10.1145/2964284.2967300.

[174] D. B. Sam, S. Surya, and R. V. Babu. Switching convolutional neural network for crowd counting. In *2017 IEEE Conference on Computer Vision and Pattern*

Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 4031–4039. IEEE Computer Society, 2017. URL: http://doi.ieeecomputersociety.org/10.1109/CVPR.2017.429, doi:10.1109/CVPR.2017.429.

[175] D. Yang, E. Yurtsever, V. Renganathan, K. A. Redmill, and Ü. Özgüner. A vision-based social distancing and critical density detection system for COVID-19. *Sensors*, 21(13):4608, 2021. URL: https://doi.org/10.3390/s21134608, doi:10.3390/s21134608.

[176] S. Abousamra, M. Hoai, D. Samaras, and C. Chen. Localization in the crowd with topological constraints. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 872–881. AAAI Press, 2021. URL: https://ojs.aaai.org/index.php/AAAI/article/view/16170.

[177] P. Torkamandi, L. Karkkainen, and J. Ott. An online method for estimating the wireless device count via privacy-preserving wi-fi fingerprinting. In O. Hohlfeld, A. Lutu, and D. Levin, editors, *Passive and Active Measurement - 22nd International Conference, PAM 2021, Virtual Event, March 29 - April 1, 2021, Proceedings*, volume 12671 of *Lecture Notes in Computer Science*, pages 406–423. Springer, 2021. URL: https://doi.org/10.1007/978-3-030-72582-2_24, doi:10.1007/978-3-030-72582-2_24.

[178] M. Vanhoef, C. Matte, M. Cunche, L. S. Cardoso, and F. Piessens. Why MAC address randomization is not enough: An analysis of wi-fi network discovery mechanisms. In X. Chen, X. Wang, and X. Huang, editors, *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security, AsiaCCS 2016, Xi'an, China, May 30 - June 3, 2016*, pages 413–424. ACM, 2016. URL: https://doi.org/10.1145/2897845.2897883, doi:10.1145/2897845.2897883.

[179] J. Martin, T. Mayberry, C. Donahue, L. Foppe, L. Brown, C. Riggins, E. C. Rye, and D. Brown. A study of MAC address randomization in mobile devices and when it fails. *Proc. Priv. Enhancing Technol.*, 2017(4):365–383, 2017. URL: https://doi.org/10.1515/popets-2017-0054, doi:10.1515/popets-2017-0054.

[180] T. D. Vo-Huu, T. D. Vo-Huu, and G. Noubir. Fingerprinting wi-fi devices using software defined radios. In M. Hollick, P. Papadimitratos, and W. Enck, editors, *Proceedings of the 9th ACM Conference on Security & Privacy in Wireless and Mobile Networks, WISEC 2016, Darmstadt, Germany, July 18-22, 2016*, pages 3–14. ACM, 2016. URL: https://doi.org/10.1145/2939918.2939936, doi:10.1145/2939918.2939936.

[181] M. Weller, J. Classen, F. Ullrich, D. Waßmann, and E. Tews. Lost and found: stopping bluetooth finders from leaking private information. In *WiSec '20: 13th*

ACM Conference on Security and Privacy in Wireless and Mobile Networks, pages 184–194. ACM, 2020. URL: https://doi.org/10.1145/3395351.3399422, doi:10.1145/3395351.3399422.

[182] A. Heinrich, M. Stute, T. Kornhuber, and M. Hollick. Who can find my devices? security and privacy of apple's crowd-sourced bluetooth location tracking system. *Proc. Priv. Enhancing Technol.*, 2021(3):227–245, 2021. URL: https://doi.org/10.2478/popets-2021-0045, doi:10.2478/popets-2021-0045.

[183] J. Martin, D. Alpuche, K. Bodeman, L. Brown, E. Fenske, L. Foppe, T. Mayberry, E. C. Rye, B. Sipes, and S. Teplov. Handoff all your privacy - A review of apple's bluetooth low energy continuity protocol. *Proc. Priv. Enhancing Technol.*, 2019(4):34–53, 2019. URL: https://doi.org/10.2478/popets-2019-0057, doi:10.2478/popets-2019-0057.

[184] J. Martin, D. Alpuche, K. Bodeman, L. Brown, E. Fenske, L. Foppe, T. Mayberry, E. C. Rye, B. Sipes, and S. Teplov. Handoff all your privacy - A review of apple's bluetooth low energy continuity protocol. *Proc. Priv. Enhancing Technol.*, 2019(4):34–53, 2019. URL: https://doi.org/10.2478/popets-2019-0057, doi:10.2478/popets-2019-0057.

[185] A. Dorais-Joncas et al. Jumping the air gap: 15 years of nation-state effort. https://www.welivesecurity.com/2021/12/01/jumping-air-gap-15-years-nation-state-effort/, 2022.

[186] M. Guri, B. Zadov, and Y. Elovici. Led-it-go: Leaking (A lot of) data from air-gapped computers via the (small) hard drive LED. In M. Polychronakis and M. Meier, editors, *Detection of Intrusions and Malware, and Vulnerability Assessment - 14th International Conference, DIMVA 2017, Bonn, Germany, July 6-7, 2017, Proceedings*, volume 10327 of *Lecture Notes in Computer Science*, pages 161–184. Springer, 2017. URL: https://doi.org/10.1007/978-3-319-60876-1_8, doi:10.1007/978-3-319-60876-1_8.

[187] M. Eichelberger, S. Tanner, G. Voirol, and R. Wattenhofer. Receiving data hidden in music. In A. Wolman and L. Zhong, editors, *Proceedings of the 20th International Workshop on Mobile Computing Systems and Applications, HotMobile 2019, Santa Cruz, CA, USA, February 27-28, 2019*, pages 33–38. ACM, 2019. URL: https://doi.org/10.1145/3301293.3302360, doi:10.1145/3301293.3302360.

[188] M. Guri, Y. A. Solewicz, A. Daidakulov, and Y. Elovici. Acoustic data exfiltration from speakerless air-gapped computers via covert hard-drive noise ('diskfiltration'). In S. N. Foley, D. Gollmann, and E. Snekkenes, editors, *Computer Security - ESORICS 2017 - 22nd European Symposium on Research in Computer Security, Oslo, Norway, September 11-15, 2017, Proceedings, Part II*, volume 10493 of *Lecture Notes in Computer Science*, pages 98–115. Springer, 2017. URL: https://doi.org/10.1007/978-3-319-66399-9_6, doi:10.1007/978-3-319-66399-9_6.

[189] J. S. Seybold. *Introduction to RF propagation*. John Wiley & Sons, 2005.

[190] G. C. Melia, M. P. Robinson, I. D. Flintoft, A. C. Marvin, and J. F. Dawson. Broadband measurement of absorption cross section of the human body in a reverberation chamber. *IEEE transactions on electromagnetic compatibility*, 55(6):1043–1050, 2013.

[191] Apple. Findmy, 2023. URL: `https://www.apple.com/icloud/find-my/`.

[192] Á. Hernández-Solana, D. P. D. Cerio, A. Valdovinos, and J. L. Valenzuela. Proposal and evaluation of BLE discovery process based on new features of bluetooth 5.0. *Sensors*, 17(9):1988, 2017. URL: `https://doi.org/10.3390/s17091988`, `doi:10.3390/s17091988`.

[193] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2980–2988. IEEE Computer Society, 2017. URL: `https://doi.org/10.1109/ICCV.2017.322`, `doi:10.1109/ICCV.2017.322`.

[194] G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

[195] R. L. Knoblauch, M. T. Pietrucha, and M. Nitzburg. Field studies of pedestrian walking speed and start-up time. *Transportation research record*, 1538(1):27–38, 1996.

[196] L. Tonetto, P. Kister, N. Mohan, and J. Ott. Ethical and privacy considerations with location based data research. *ArXiv*, 2024.

Publication 1

©2018 IEEE, reprinted with permission from:

B. Alipour, L. Tonetto, A. Y. Ding, R. Ketabi, J. Ott and A. Helmy, "Flutes vs. Cellos: Analyzing Mobility-Traffic Correlations in Large WLAN Traces," IEEE INFOCOM 2018 - IEEE Conference on Computer Communications, Honolulu, HI, USA, 2018, pp. 1637-1645, doi: 10.1109/INFOCOM.2018.8486360.

RightsLink

Sign in/Register

Flutes vs. Cellos: Analyzing Mobility-Traffic Correlations in Large WLAN Traces

Conference Proceedings:
IEEE INFOCOM 2018 - IEEE Conference on Computer Communications

Author: Babak Alipour

Publisher: IEEE

Date: April 2018

Copyright © 2018, IEEE

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK CLOSE WINDOW

© 2024 Copyright - All Rights Reserved | Copyright Clearance Center, Inc. | Privacy statement | Data Security and Privacy | For California Residents | Terms and ConditionsComments? We would like to hear from you. E-mail us at customercare@copyright.com

Title: *Flutes* vs. *Cellos*: Analyzing Mobility-Traffic Correlations in Large WLAN Traces
Authors: Babak Alipour (UFL), **Leonardo Tonetto** (TUM), Aaron Yi Ding (TUM), Roozbeh Ketabi (UFL), Jörg Ott (TUM), Ahmed Helmy (UFL)
Venue: 2018 IEEE Conference on Computer Communications, INFOCOM 2018, Honolulu, HI, USA
Publishing date: April 16-19, 2018
Reference: [2]

Publication Summary

In this article, we focus on understanding the interplay between mobility and traffic patterns in mobile networks. Previous models have mostly considered mobility and traffic in isolation, without capturing their interdependencies. Our study aims to quantitatively characterize the mobility and traffic patterns for laptops (referred to as "cellos") and smartphones (referred to as "flutes") and analyze their differences and relationships.

We conducted a multi-dimensional analysis using large datasets capturing mobility and traffic features. The datasets included WLAN Access Point (AP) logs and NetFlow records, totaling over 30TB of data and covering 300,000 devices. We developed a systematic framework called "FLAMeS" (Framework for Large-scale Analysis of Mobile Societies) to analyze the data.

The main contributions of the study include integrated mobility-traffic analyses, a comparison between flutes and cellos, and the development of the FLAMeS framework. The integrated analysis provides insights into the correlations between various mobility and traffic features, identifying gaps in existing models and paving the way for future work. The comparison between flutes and cellos highlights the differences in mobility and traffic characteristics between laptops and smartphones. The FLAMeS framework enables the systematic analysis of large-scale datasets and can be applied to other studies involving multi-sourced data.

Our experimental setup involved collecting datasets from WLAN AP logs and NetFlow records from a university campus. The datasets were preprocessed and used to classify devices into flutes and cellos. And the analysis was conducted to investigate spatio-temporal patterns and correlations between mobility and traffic features.

Overall, the study provides valuable insights into the interplay between mobility and traffic in mobile networks, contributing to the development of integrated mobility-traffic models and enhancing the understanding of user behavior and network performance.

Contributions

The idea for the paper originated from the collaborative work I did with Babak Alipour, with the supervision of Jörg Ott and Ahmed Helmy. We cleaned, processed and analyzed the data. I designed and executed the mobility part of the study, including analyses, plots, figures and tables, as well as part of the analysis and observations of the combined

mobility traffic analysis. I wrote and reviewed all text related to mobility as well as conclusions of our overall findings. All authors reviewed the text.

Flutes vs. *Cellos*: Analyzing Mobility-Traffic Correlations in Large WLAN Traces

Babak Alipour* babak.ap@ufl.edu
Leonardo Tonetto† tonetto@in.tum.de
Aaron Yi Ding† ding@in.tum.de
Roozbeh Ketabi* roozbeh@ufl.edu
Jörg Ott† ott@in.tum.de
Ahmed Helmy* helmy@ufl.edu

*Computer and Information Science and Engineering
University of Florida, Gainesville, USA

†Department of Informatics
Technical University of Munich, Munich, Germany

***Abstract*—Two major factors affecting mobile network performance are *mobility* and *traffic* patterns. Simulations and analytical-based performance evaluations rely on models to approximate factors affecting the network. Hence, the understanding of mobility and traffic is imperative to the effective evaluation and efficient design of future mobile networks. Current models target either mobility or traffic, but do not capture their interplay. Many trace-based mobility models have largely used pre-smartphone datasets (e.g., AP-logs), or much coarser granularity (e.g., cell-towers) traces. This raises questions regarding the relevance of existing models, and motivates our study to revisit this area. In this study, we conduct a multi-dimensional analysis, to *quantitatively* characterize mobility and traffic spatio-temporal patterns, for laptops and smartphones, leading to a detailed integrated mobility-traffic analysis. Our study is *data-driven*, as we collect and mine capacious datasets (with 30TB, 300k devices) that capture all of these dimensions. The investigation is performed using our systematic (*FLAMeS*) framework. Overall, dozens of mobility and traffic features have been analyzed. The insights and lessons learnt serve as guidelines and a first step towards future *integrated mobility-traffic models*. In addition, our work acts as a stepping-stone towards a richer, more-realistic suite of *mobile test scenarios* and *benchmarks*.**

I. Introduction

Human mobility has been studied extensively and many models have been derived. The spectrum ranges from simple synthetic mobility models to complex trace-based models, capturing different properties with varying degrees of accuracy [1], [2]. Similarly, network traffic has been studied increasingly for wireless networks: for rather stationary users (as in WLANs) (e.g., [3], [4]) and potentially mobile users as for cellular networks (e.g., [5], [6]). Such analyses range from metrics such as flow count, sizes, and traffic volume to service usage (e.g., visited web sites, backend services).

Both mobility and network usage, characterize different aspects of human behavior. In this sense, we have a *mobility plane* and a *(network) traffic plane*. In reality, these two planes are likely interdependent. Human mobility may be influenced by network activity; for example, a person slowing down to read incoming messages. Also, network activity may be influenced by mobility and location; stationary users may produce/consume more data than those walking, and people may use different services in different places [7].

In earlier studies, this interdependence has not been widely considered, and models for both mobility and network traffic planes have been developed and evaluated largely in isolation. For example, when evaluating mobile systems' performance, traffic generation generally follows regular patterns, drawn from common simple distributions (e.g., exponential or uniform), while assuming neither transmission nor reception of data impacts mobility. Simply observing people walking while staring at (or reacting to) their smartphones suggests, however, that such interdependencies need to be captured properly. Understanding the mobility-traffic interplay is imperative to the effective evaluation and efficient design of future mobile algorithms ranging from user behavior prediction and caching, to network load estimation and resource allocation.

In this paper, we take a stab at understanding the interconnection of the mobility and traffic planes. To do this properly, we need to consider the nature of mobile devices people use: one class of devices is merely intended for stationary use, typically while the user is seated—this primarily holds for laptop computers, dubbed *cellos*. In contrast, another class—smartphones, which we refer to as *flutes*—lend themselves to truly mobile use[1]. We focus our analysis on these two classes because they have been around long enough to have extensive datasets to build upon. We stipulate that the interconnection of the mobility and traffic is modulated by the device(s) a mobile user is carrying. Therefore, we follow two main lines of investigation: we develop a framework to differentiate between cellos and flutes, and study both the mobility and traffic patterns for each of those types.

Specifically, the main goal of this paper is to quantitatively investigate the following questions in-depth: *(I) How different are mobility and traffic characteristics across device types, time and space? (II) What are the relationships between these characteristics? (III) Should new models be devised to capture these differences? And, if so, how?*

To answer these questions, a multi-dimensional (comparative) analysis approach is adopted to investigate mobility and traffic spatio-temporal patterns for flutes and cellos. We drive our study with capacious datasets (30TB+) that capture all the above dimensions in a campus society, including over 300k devices (Sec. IV). A systematic *F*ramework for *L*arge-scale

[1]Throughout this paper we shall use the terms *flutes* and *smartphones* interchangeably, and the terms *cellos* and *laptops* interchangeably.

*A*nalysis of *M*obil*e* *S*ocieties (*FLAMeS*) is devised for this study, that can also be used to analyze other multi-sourced data in future studies. Our main contributions include:

1) *Integrated mobility-traffic analyses* (Sec. VII): this study is the first to quantify the correlations of numerous features of mobility and traffic simultaneously. This can identify gaps in existing mobile networking models, and reopen the door for future impactful work in this area.
2) *Flutes vs. Cellos analysis* (Sec. V–VI): the device-type classification presented here, facilitates another important dimension to understand. This is particularly important as new generations of portable devices are introduced, that are different than the laptops, traditionally considered in earlier studies.
3) *Systematic multi-dimensional investigation framework* (Sec. III): *FLAMeS* provides the scaffolding needed to process, in multiple dimensions, many features of large sets of measurements from wireless networks, including AP-logs and NetFlow traces. This systematic method can apply to other datasets in future studies.

II. Related Work

To characterize mobility and network usage, existing studies have covered various aspects, including human mobility, device variation, and dataset analysis.

Human Mobility: Given its importance in various research areas, human mobility has received significant attention. We refer the reader to [1], [2] for surveys of mobility modeling and analysis. For spatial-temporal patterns, [8] and [9] reveal the regularity and bounds for predicting human mobility using cellular logs. A recent study has highlighted the importance of combining different datasets to study various features simultaneously [10]. Our observations are similar to [8], [9] which reassure the intrinsic properties of human mobility, despite the differences in granularity and population characteristics across datasets. To advance the understanding of human mobility, we integrated different datasets to correlate mobility and network traffic.

Device Variation: Usage and traffic patterns of different device types have been studied from various perspectives ([11], [12], [13], [14], [15], [16]). However, those findings are based on classifications that rely on either MAC addresses or HTTP headers solely. The former is rather limited and the latter may have serious privacy implications and are often unavailable. In [17], authors use packet-level traces from 10 phones and application-level monitoring from 33 Android devices to analyze smartphone traffic. Although this allowed fine-grained measurements, the approach is invasive and limited in scalability, leading to small sample sizes and restricted conclusions. They also do not compare the traffic of smartphones with that of "stop-to-use" wireless devices (i.e. cellos) nor do they measure spatial metrics. To characterize usage pattern for users with multiple wireless devices, Das et al. carried out wireless trace analysis on a university campus, covering more than 32k users [18]. Their study revealed usage difference between laptops/tablets and smartphones in terms of time, packets, content, intermittent and overlapping usage. Compared with our study to correlate mobility and traffic, their work targets at device usage patterns and security aspects. In our method, the combination of MAC and NetFlow allowed us to classify majority of observed devices while preserving users' privacy.

Dataset Analysis: The most recent work on WLAN traces [19] revealed surprising patterns on increases of long-term mobility entropy by age, and the impact of academic majors on students' long-term mobility entropy. The authors of [7] investigated correlations and characteristics of web domains accessed by users and their locations of users based on NetFlow and DHCP logs from a university campus in 2004. They propose a simulation paradigm whose parameters are extracted, producing realistic scenarios for simulations. However, the study uses data from pre-smartphone era and does not distinguish between device types. It also does not analyze the relationship between mobility and traffic. On both WiFi and cellular networks, the authors of [20] performed an in-depth study on smartphones traffic, highlighting the benefits and caveats/limitations of using MPTCP. Distributions of flow, Inter-arrival time (IAT) and arrival rate at APs of "static" flows has been analyzed against popular distributions (e.g., Exp, Weibull, Pareto, Lognormal) [21]. Lognormal was found to best fit the flow sizes, while at small time scales (i.e. hourly), IAT was best described by Weibull but parameters vary from hour to hour. We analyze flows on a much larger scale, newer dataset including smartphones, and identify Lognormal distribution as the best fit for flow sizes, and beta as best for IAT, regardless of device type. Xu et al. conducted a large scale ISP trace analysis that covers over 9600 cellular towers and 150k subscribers of city Shanghai in August 2014 [22]. Their study identified the mapping between time-domain traffic patterns and five types of urban functional regions, yielding several insights on mobile traffic patterns across time, location and frequency domains. This work is complementary to our study on campus WLAN traces as they focus on cellular networks in urban areas.

III. Systematic multi-dimensional analysis

To methodically analyze statistical characteristics and correlations in multiple dimensions, we introduce the *FLAMeS* framework (Fig. 1). The main components include: I. Data collection and pre-processing, II. Flutes vs. cellos mobility and traffic analysis, and III. Integrated mobility-traffic analysis.

The two main purposes of this work are to understand and **quantify** the *gaps between flutes and cellos*, and the *interaction between the mobility and traffic dimensions*. Individual mobility and traffic analyses for flutes and cellos are conducted in Sections V and VI, respectively, with detailed reporting for spatio-temporal features showing significant gaps. In Sec. VII, the most important mobility and traffic features are identified and their correlation quantified.

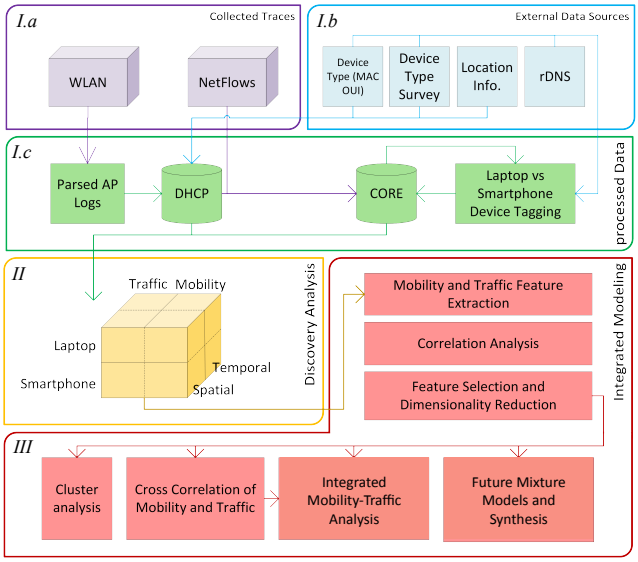


Fig. 1: *FLAMeS* system overview.

IV. Experimental Setup and Datasets

We drive our framework with large-scale datasets from multiple sources, capturing the mobility and traffic features in different dimensions. In this section, we introduce the two data sets and their preprocessing, and present the device type classification into flutes and cellos.

The *input datasets* in this study are specifically chosen to capture: 1. location, mobility and network traffic information, 2. smartphone and laptop devices, 3. spatio-temporal features, and 4. scale in number of devices and records. The total size is >**30TB**, consisting of two main parts: WLAN Access Point (AP) logs, and Netflow records (details in Tables I, II).

A. WLAN AP Logs

These logs are collected from 1760 APs in 138 buildings over 479 days on a university campus, and contain association and authentication events from 316k devices in 2011-2012. It contains over 555M records, with each record including the device's MAC and assigned IP addresses, the associated AP and a timestamp. Locations of the APs are approximated by the building locations where they are installed, i.e., (longitude, latitude) of Google Maps API. To validate this, we fetched 8000 mapped APs around the campus area from a crowd-sourced service, *wigle.net*. For the 130 matched APs in 42% of buildings (i.e., 58 bldgs), all were less than 200m from their mapped location; an error of less than 1.5% of the campus area. This is very reasonable for our study purposes.

B. NetFlow Logs

Over **76 billion** records of NetFlow traces were collected from the same network, over 25 days in April 2012. A *flow* is defined as a consecutive sequence of packets with the same transport protocol, source/destination IP and port number, as identified by the collecting gateway router. An example of major Netflow data fields is presented in Table II.

The NetFlow records are matched with the wireless associations (from the AP logs) using the dynamic MAC-to-IP address mapping from the DHCP logs. We refer to the result as *CORE* dataset (Table I). They are also augmented with location and website information using reverse DNS (rDNS)[2].

TABLE I: Summary of datasets. B=billion.

	# Records		Traffic Vol. (TB)		# MAC	
	DHCP	CORE	TCP	UDP	WLAN	CORE
Flutes	412.0 M	2.13 B	56.18	4.50	186.0 K	50.3 K
Cellos	101.0 M	4.20 B	73.85	12.90	93.2 K	27.1 K
Total	557.5 M	6.53 B	134.39	17.61	316.0 K	80.0 K

C. Device Type Classification

To classify devices into flutes and cellos, we utilize several observations and heuristics. To start, note that a device manufacturer (with OUI) can be identified based on the first 3 octets of the MAC address[3]. Most manufacturers produce one type of device (either laptop or phone), but some produce both (e.g., Apple). In the latter case, OUI used for one device type is not used for another. We conducted a survey to help classify 30 MAC prefixes accurately. Using OUI and survey information, we identify and label 46% of the total devices (90k cellos and 56k flutes). Then, from the NetFlow logs of these labeled devices, we observe over 3k devices (92% of which are flutes) contacting *admob.com*; an ad platform serving mainly smartphones and tablets (i.e. flutes). This enables further classification of the remaining MAC addresses. Finally, we apply the following heuristic to the dataset: (1) obtain all OUIs (MAC prefix) that contacted *admob.com*; (2) if it is unlabeled, mark it as a flute. Overall, over 270k devices were labeled (180k as flutes), covering 86% of the devices in AP logs and 97% in NetFlow traces, a reasonable coverage for our purposes. Out of $\approx 80k$ devices in the NetFlow logs, $\approx$ 50K are flutes and $\approx$ 27K cellos.

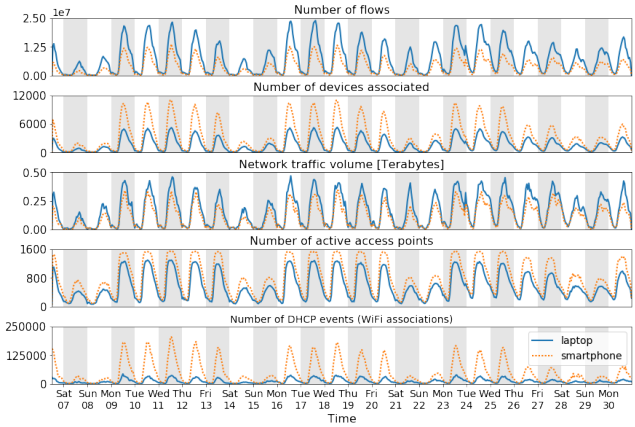


Fig. 2: Time series for 25 days of combined AP-NetFlow Core traces

Fig. 2 shows the temporal plot for the combined traces over 25 days, after device classification. Throughout, the number

[2]Detailed dataset merging and the query computing system are available in the Appendix II of our technical report [23].

[3]MAC address randomization does not affect our association trace.

of flows and total traffic volume is clearly higher for cellos, even with an overall higher number of flutes connected. Also note the device activities in a *diurnal* and *weekly* cycles, with the peaks occurring during weekdays, as expected. Wed, 25th, was the last day of classes, explaining the decline in network activity afterwards. The plot motivates our analyses for flute vs cellos, over *weekends vs weekdays*, for the rest of the study.

V. Mobility Analysis

This section covers the *temporal* and *spatial* mobility analyses. For all metrics, unless otherwise noted, we investigate 479 days. A summary of studied metrics and their most significant statistical values are presented in Tab. III along with mean and median ratios for comparison. From that list, we further investigate in this section those metrics that show the most interesting or non-trivial differences between *flutes* and *cellos*.

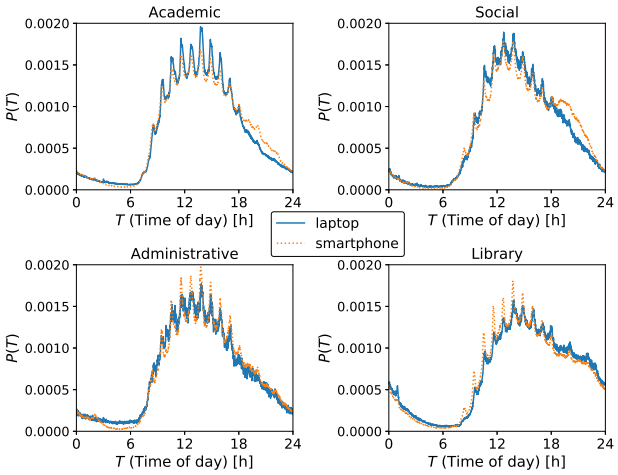


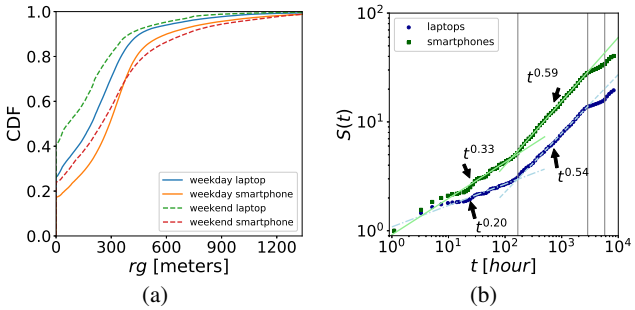
Fig. 3: PDF Session start over time of the day.

A. Session start probability

We define a session as the period between WLAN associations. The distributions of session start times across the day for four building categories are depicted in Fig. 3. aspects of analysis for this metric. The start times of the Sessions match the periodic beginning of classes, but mainly in *Academic* buildings, where users move mostly at the start and end of classes. In these places, activity drops sharply for *cellos* at 5pm, with considerable *flutes* activity until 8pm. For *Social* and *Library* buildings, *the probability of new sessions remains higher for a few more hours into the evening, and the times users tend to leave are more spread out*. We do not make similar observation during weekends, which is expected when the day is, unlike weekdays, not governed by a class schedule. For most visitors, the session start distributions show a smooth shape and no significant differences between device types (omitted for brevity).

B. Radius of Gyration

This metric, GYR, captures the size of the geospatial dispersion of a device's movements, denoted by r_g and computed



(a) (b)

Fig. 4: (a) Radius of gyration (rg for the device types). (b) Visited locations $S(t)$. Vertical lines at 7, 120 and 240 days.

as $r_g = \frac{1}{N}\sum_{k=1}^{N}(\vec{r_k} - \vec{r_s})^2$, where $\vec{r_1}, ..., \vec{r_N}$ are positional vectors of a device and $\vec{r_s}$ is its center of gravity.

Grouping devices by their r_g after six months of observation, we look at its evolution since the first time they are observed. Unsurprisingly (cf. [8]), after an initial transient period of about one week, this value stabilizes even across different semesters (not shown).

We split the traces into weekdays and weekends, presenting the distributions in Fig. 4a. Comparing these groups for *cellos*, we notice a substantial reduction in their overall mobility while, for *flutes*, this difference is not so pronounced. This might be due to students having fewer activities on weekends, a tendency to study at a single building like a library, or just not carry their cellos; we will revisit this aspect in Sec. VII. *Flutes*, being "always-on" devices, are able to capture movements at pass-by locations, dining areas, and bus stops and thus are better suited to capture the fine-granular mobility of their users than cellos.

Despite the 8.1km^2 area of the campus (approximate radius of 1.42km), buildings with related fields of study (e.g. Fine Arts) are somewhat clustered. Computing the distance between the k-nearest neighboring buildings, for $k = 22$ and $k = 9$ (average number of visited buildings for *flutes* and *cellos*) the median distances are 295m and 172m, respectively. Due to their focus on classes, students attending have limited area of activity on weekdays, which explains the observed *radius of gyration*.

We also evaluated: (1) *diameter* DIA, the longest distance between any pair of $\vec{r_k}$ points; (2) *max jump* LJM, the longest distance between a pair of consecutive $\vec{r_k}$ points; and (3) *total trajectory length* TJM, the sum of all trips made by a device. The distributions of these metrics are similar to *Radius of Gyration* and therefore not shown. Table III summarizes the most significant statistical values for these metrics.

C. Visitation preferences and interests

We count the number of unique buildings visited by a user, BLD, and define a *preferred building* as the location where a device has spent most of its time in a given day, which we measure in minutes and refer to as PDT. We approximate the latter by the formula $t_b = \sum_{k=1}^{N_b} S_k$, where t_b is the time spent, N_b the total number of sessions and $S_1 ... S_N$ the time

TABLE II: NetFlow (top) and AP logs/DHCP (bottom) sample data

Start time	Finish time	Duration	Source IP	Destination IP	Protocol	Source port	Destination port	Packet count	Flow size
1334332274.912	1334332276.576	1.664	173.194.37.7	10.15.225.126	TCP	80	60482	157	217708

User IP	User MAC	AP name	AP MAC	Lease begin time	Lease end time
10.130.90.3	00:11:22:33:44:55	b422r143-win-1	00:1d:e5:8f:1b:30	1333238737	1333238741

TABLE III: General results for mobility. Upper values are for weekdays and lower ones for weekends (in red color). **LJM**: maximum jump [m]; **DIA**: diameter [m]; **TJM**: total trajectory length [m]; **GYR**: radius of gyration [m]; **BLD**: no. uniq. buildings; **APC**: access point count; **PDT**: time spent at preferred building [minutes]; **DLT**: total session time at each building.

	Flutes (F)			**Cellos (C)**			**Ratio (C/F)**	
	μ	***mdn***	σ	μ	***mdn***	σ	μ	***mdn***
LJM	435	296	813	178	1	624	0.409	**0.003**
	350	168	683	97	1	312	0.277	0.006
DIA	549	411	874	195	1	642	0.355	**0.002**
	425	179	739	107	1	338	0.252	0.006
TJM	1582	707	2336	378	1	1444	0.239	**0.001**
	1036	279	1793	252	1	1766	0.243	0.004
GYR	396	290	2725	321	191	3265	1.102	1.019
	330	248	1368	178	65.1	1800	1.247	1.4
BLD	5.4	3	5.6	1.8	1	2.1	0.811	0.659
	2.8	2	4.1	1.5	1	1.8	0.539	0.262
APC	11.8	6	13.3	3.7	2	4.8	0.333	0.333
	7.2	4	8.8	3	2	3.8	0.536	0.5
PDT	225	161	219	248	164	254	0.314	0.333
	223	135	272	278	189	292	0.417	0.5
DTL	316	235	302	316	217	305	1	0.92
	326	247	308	316	221	309	0.97	0.89

duration of *each session* at a building b, here referred as DLT. Interestingly, cellos have slightly longer stays but both have medians around 2:40 hours. The similarity of the distributions, combined with a lower number of visited locations indicate that cellos are used mostly when users remain longer periods at places.

Fig. 4b highlights the differences between *flutes* and *cellos* on the required time t to visit $S(t)$ locations. *After an initial exploration period of one week the rates of new visits change similarly for both device types, and new exploration rates show up at 120 and 240 days.* These could be explained by the weekly schedules of the university as well as the usual length of a lecture term (≈ 4 months).

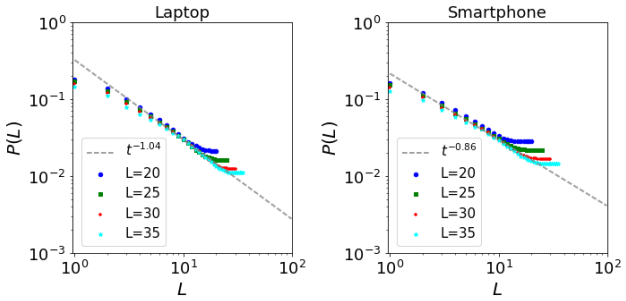


Fig. 5: Zipf's plot on L visited access points.

We also consider the number of unique access points a device associates with, APC, which provides a finer spatial resolution than the building level. Furthermore, the probability of finding a device at its *L-th* most visited access point is shown in Fig. 5. When taking buildings as aggregating points for location, the values become $L^{-1.36}$ for *cellos* and $L^{-1.16}$ for *flutes*. These approximations validate previous work on human mobility [8], yet highlighting differences between device types.

D. Sessions per building

To study AP utilization over time, we look at the session duration distribution, or session duration dispersal kernel P(t), depicted in Fig. 6. The smaller inner plots represent the same metric, limited to four types of buildings.

We noted that the five-minute spikes correspond to default idle-timeout for the used WiFi routers. On the other hand, the *knees* at 1 and 2 hours could be explained by the typical duration of classes. They are only noticeable at Academic buildings (shown inside inner plots) and during weekdays (not shown). This leads us to conclude that despite the differences in distributions of device types, *flutes* and *cellos* present *certain similarities in their usage, such as during classes*. To differentiate *pass-by* access points, we examine all sequences of three unique APs where all session durations are lower than 5 minutes (typical idle-timeout). We observed these APs clustered at buildings that also had major bus stops nearby.

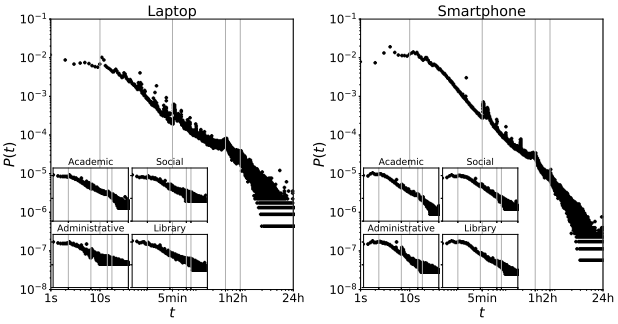


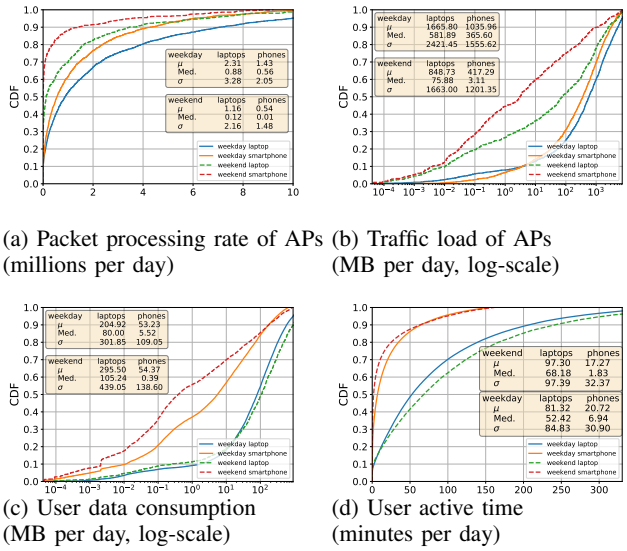
Fig. 6: Probability $P(t)$ of session duration t.

VI. Traffic Analysis

In this section, we compare different *traffic* characteristics, across *device types*, *time* and *space*. For this purpose, we start with statistical characterization of *individual* flute and cello flows. Next, we measure how these flows, *put together*, affect the network patterns across APs and buildings. Finally, *user behavior* is analyzed by monitoring weekly cycles, data rates, and active durations. By quantifying *temporal* and *spatial* variations of traffic across device types, we make a case for new models to capture such variations based on the most relevant attributes. Table IV summarizes the results.

A. Flow-level statistical characterization

We compare the following distributions using maximum likelihood estimation (MLE) and maximizing goodness-of-fit



(a) Packet processing rate of APs (millions per day)

(b) Traffic load of APs (MB per day, log-scale)

(c) User data consumption (MB per day, log-scale)

(d) User active time (minutes per day)

Fig. 7: Distribution plots

estimation: Gaussian, Exponential, Gamma, Weibull, Logistic, Beta and Lognormal[4].

1) Size: Flow size is the sum of bytes for all packets within a single flow. On weekdays, average *size of individual flute flows* is $> 2x$ **larger** *than cello flows* (2070 vs. 822 bytes), while median is $> 4x$ **larger** (678 vs. 142 bytes). There are no significant changes on weekends (See Appendix IV in [23] for details).

The average packet size within a single flow also provides insight into packet-level behavior of services on mobile devices. We notice that the average *packet size of flute flows* is ≈**50% larger** *than that of cellos* (212 vs 144 bytes on weekdays, 205 vs 142 on weekends). Comparing weekdays and weekends, median size of flute packets drops on weekends whereas it remains *the same* for cellos. In fact, comparing cello flows on weekdays and weekends shows *no significant difference* in terms of average packet size (p-value$>$ .05). In spite of smaller flows, the average cello generates *2.7 times traffic* as an average flute because the average cello is responsible for *3.7 as many flows* as a flute. Analyzing distributions of flow size and average packet size in our datasets shows that *Lognormal* distribution is the best fit, with varying parameters for each device type (See Appendix IV in [23] for details).

2) Packets: This metric is the count of packets within each flow. The mean and median packet counts per flow are 7.06 and 5 in flutes and 3.64 and 2 in cellos, during weekdays. The means drop slightly on weekends. Packet counts per flow match the *Lognormal* distribution well for flows of both device types. The average flute flow is bigger in size and has *more packets* (with higher variance) but there are *fewer* flows coming from these devices. This is analyzed further for TCP/UDP flows (Sec. VI-A5).

3) Runtime: Flow runtime is the period of time the flow was active (equal to a flow's $finishtime - starttime$). Flute flows have a mean and median of 1868ms and 128ms respectively on weekdays, while these numbers are 1639ms and 64ms for cellos. Both device types show increase in means during weekends (flutes by 204 and cellos by 164), indicating that although there are fewer devices online during weekends, they are more active. The low medians in either group corresponds to many *short-lived* flows with few packets, showing little variation across device type, time or space.

4) Inter-arrival times (IAT): Median of the flow *IAT*[5] at access points is 6ms for cello flows and 4ms in case of flutes, on weekdays (similar on weekends), which suggests that the majority of access points handle flows from either device type at nearly the same rate. However, average *IAT* is ≈ 143ms for flute and ≈ 78ms for cello flows, as there are more cellos with very high rate of flows. Flow *IAT* in our datasets matches a **beta** distribution well (See Appendix IV in [23]) with a *very high estimated kurtosis* and *skewness* (estimated at 58 & 6.9 respectively). The high estimated kurtosis illustrates that there are *infrequent extreme values*, which explains the observed highly elevated standard deviation of *IAT*. Higher average IAT of flutes, combined with the higher standard deviation compared to cellos (596 vs 284), shows that *flutes face more extreme periods of inactivity, which can be caused by higher mobility and packet loss*.

5) Protocols: TCP accounts for *78.5%* of cello flows (**84.6% of bytes**) and *98.2%* of flute flows (**91.6% of bytes**). The higher presence of UDP in cellos is reasonable, considering that UDP applications (e.g., multi-player games, video conferencing and file sharing) are more likely to be used with cellos. Comparing the number of packets in flows, in case of TCP, the average number of packets in cello flows is almost half that of a flute flow (4.6 vs 8.8), and the average packet size of flutes is 22% higher than that of cellos. This supports our earlier observation regarding the bigger flows sizes of flutes. However, for UDP, the two device types are similar in terms of average packet count per flow (2.5 & 2.87 for cellos and flutes respectively) and average packet size (119 for both). This conforms to low latency requirements of many UDP applications.

Given these differences, traffic classification using machine learning [26] could benefit from considering device types to train models. We investigate this in VII-B.

After establishing the similarities and differences of flows, the next step is to evaluate whether the individual variations in flows lead to different *aggregate traffic behaviors* from viewpoint of the network.

[4]For distribution comparison, significance threshold $p - value$ is set at .05.

[5]IAT is important in simulation and modeling of networking protocols, traffic classification [24], congestion control and traffic performance [25]. Our flow-level IAT analysis can also be used for measuring delay and jitter effects.

B. Network-centric (spatial) analysis

We examined the load of APs in all buildings on a daily basis to provide insight into differences from the viewpoint of the network. For each AP, we calculate flow metrics for every weekday and weekend.We focus our analysis on the first three weeks of NetFlow traces to avoid significant user behavior change during exams period, as already shown in Fig. 2.

First, we measure the daily packet and flow arrival rates at APs. The median flow rates are $42k$ and $20k$ per weekday for cellos and flutes respectively (7.5k and 0.5k on weekends). The average number of cello packets processed daily by APs is **1.6 times higher** than flute packets (Fig. 7a). Each AP handles, on average during weekdays, ≈ 27 *cello packets per second* and ≈ 17 *flute packets per second*, dropping to ≈ 13.5 and ≈ 6.25 on weekends. This indicates that, during the weekends, a high percentage of access points are not utilized, with *60% of APs seeing no flute flows* and *70% receiving no cello flows*. However, *at least one AP in >80% of buildings sees traffic*, supporting observations of less mobility during weekends.

Next, we look at traffic volume. On average weekdays, 90% of APs handle < $5GB$ of cello traffic ($2.5GB$ on weekends), whereas the same percentage handles < $3GB$ of flute traffic ($1GB$ on weekends) (Fig. 7b). Flutes are more mobile, visit a higher number of unique APs and have bigger flow sizes but they are still responsible for *less overall network load*.

Thus, the individual differences of flute and cello flows result in *heterogeneous aggregate traffic patterns* in time (different days) and space (APs at different buildings)[6]. With that established, in order to take steps towards modeling and simulation, we also need to analyze the behavior of users.

C. User behavior (temporal) analysis

Here, we measure traffic patterns from a user-centric perspective. We identified gaps in diurnal and weekly cycles (Fig. 2) as well as traffic flow features of individual *users* including data consumption, packet rates, and network activity duration.

1) Data consumption: Fig. 7c shows daily data consumption, with 90% of cellos consuming **< 700MB** and 90% of flutes using **< 200M** on weekdays. Surprisingly, for cellos on campus during weekends, average data consumption is even higher whereas data consumption of flutes drops sharply.

2) Packet rate: On weekdays, cellos on average generate **≈318K packets**, while flutes only average **≈84K packets** per day. On weekends, the few on-campus cellos see greatly increased number of packets, with an average daily packet rate of ≈ 495K. Weekend flutes also have a modestly *increased* packet count, with an average of ≈ 96K flows.

3) Active duration: Total active time of devices serves well to demonstrate the differences between time spent online by users of different device types. We rely on NetFlow to measure **'active'** time instead of AP association time. This allows us to distinguish user's *idle* presence in the network from its *activity* periods. Cellos have **4x** average active time compared to flutes in our traces (≈ 81 vs ≈ 21 min on weekdays, ≈ 97 vs ≈ 17 min on weekends). Overall, 90% of cellos are active for <**3.5h** and 90% of flutes are active for <**1h** (Fig. 7d). As evident in various metrics, the cellos appearing on weekends are more active than the average cello on weekdays.

Overall, the data consumption of flutes seems to be *more bursty* in nature, with **bigger** flows and **lower active duration**. This could be due to more intermittent usage of flutes and also bundling of network requests to save battery on these devices. In addition, there are fewer devices on campus during weekends, but those remaining devices are more active and consume more data than average.

VII. Integrated Mobility-traffic analysis

By studying the relationship between features from mobility and network traffic, we examine whether the *fusion* of these dimensions provides a case for the necessity of an *integrated mobility-traffic model* and introduce steps towards a combined model (Sec. VII-B).

A. Feature engineering

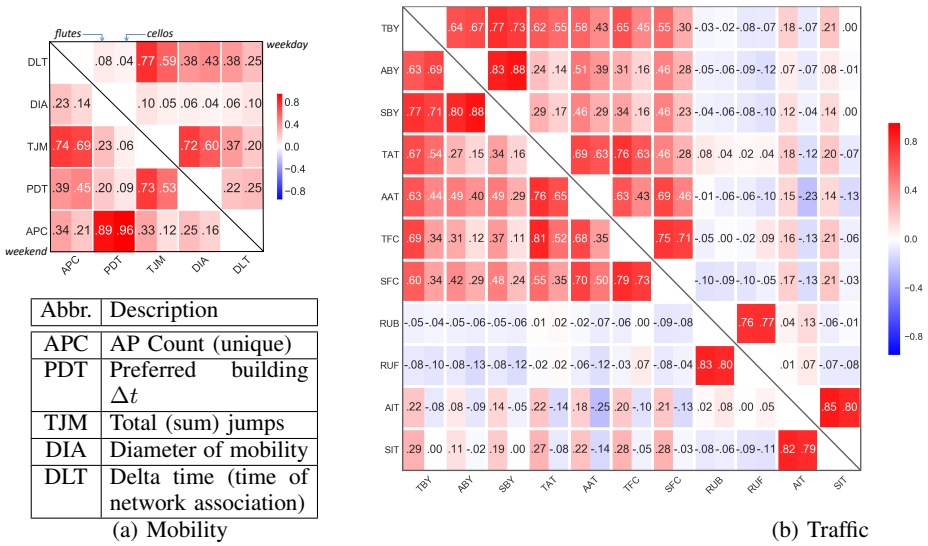
To simplify analysis and interpretation, and reduce dimensionality, we identify the most important features. First, we study the relationships among variables from *mobility* and *traffic* dimensions separately. Then, from this subset of combined features, we investigate whether clusters of user devices appear in the dataset. For this, we use correlation feature selection (*CFS* [27]), to obtain uncorrelated features[7], but highly correlated to the classification. Finally, we quantify correlations between mobility and traffic metrics (See abbreviations in Fig. 8).

1) Mobility: The *CFS* algorithm was run on 8 features (in Sec. V), and kept only *5* (to be used in the cross-dimension analysis). Fig. 8a visualizes the linear dependence between mobility features, comparing flutes and cellos on weekdays and weekends. Close inspection reveals temporal correlation relationships. For example, for cellos on weekdays, there is a **strong** correlation (0.96) between preferred building time (*PDT*) and time of network association (*DLT*), but weak correlation (0.1) on weekends, suggesting that most of weekend online time is spent at preferred buildings (e.g., libraries).

2) Traffic: We extract statistical measures for traffic metrics (Sec. VI) per device per day. The *CFS* algorithm was run on 19 features, reducing them to 11. A summary of these metrics is provided in Table IV. The correlations are depicted in Fig. 8b. The analysis shows us that average number of packets and bytes are positively correlated, but negatively correlated with variance of bytes and uncorrelated with IAT. Average IAT (*AIT*) seems to be mostly independent from other traffic features, but as *AIT* increases, its standard deviation (*SIT*) also greatly increases which could be due to device mobility; bearing further investigation on traffic-mobility interactions. Interestingly, active time is *weakly correlated* with number of flows and packets, which shows that users who remain online longer are *not* necessarily consuming traffic at a high rate.

[6]A more in-depth analysis is presented in the Appendix IV in [23].

[7]We show Pearson correlation for simplicity, and omit non-linear correlation for brevity.



Abbr.	Description
APC	AP Count (unique)
PDT	Preferred building Δt
TJM	Total (sum) jumps
DIA	Diameter of mobility
DLT	Delta time (time of network association)

(a) Mobility

Abbr.	Description
TBY	Total flow bytes
ABY	Avg. flow bytes
SBY	Std. flow bytes
TAT	Total active time
AAT	Avg. active time
TFC	Total flow count
SFC	Std. flow counts
RUB	UDP bytes / total bytes
RUF	UDP flows / total flows
AIT	Avg. IAT
SIT	Std. IAT

(b) Traffic

Fig. 8: Correlation plots for (a) *mobility* and (b) *traffic* features. Each cell's left half is for flutes and right half is for cellos, the upper right triangle is for weekdays and the lower left for weekends.

TABLE IV: Summary of traffic features used for integrated mobility-traffic analysis (per device, averaged over all days; see Fig. 8 for abbreviations). Upper values are for weekdays and lower ones for weekends (in red color).

	Flutes (F)			Cellos (C)			Ratio (C/F)	
	μ	***mdn***	σ	μ	***mdn***	σ	μ	***mdn***
TBY	96.77	11.47	194.52	373.08	144.68	554.54	3.85	**12.61**
[MB]	80.96	0.86	195.15	448.87	180.23	623.86	5.54	209.56
ABY	5.48	0.74	14.02	15.67	7.34	25.81	2.85	**9.91**
	4.54	0.15	14.16	18.06	8.34	28.71	3.97	55.6
SBY	10.56	1.57	23.76	30.59	13.77	49.82	2.89	**8.77**
	8.09	0.13	21.48	33.21	15.42	53.39	4.10	118.61
TAT	1,330	388.6	2,517	5,123	3,003	6,444	3.85	7.73
	1,059	90.89	2,497	5,883	3,861	6,934	5.55	42.48
AAT	63.14	27.97	86.69	188.26	166.93	138.70	2.98	5.96
	50.60	12.98	85.27	206.89	184.17	156.53	4.08	14.18
TFC	7.2	1.7	15.61	33.5	17.1	60.10	4.65	**10.05**
[K]	5.7	0.3	15.01	38.5	20.6	88.52	6.75	68.66
SFC	515.6	177.3	907.7	1,640	1,181	2,081	3.18	6.66
	361.05	30.18	796.6	1,673	1,215	2,098	4.63	40.27
RUB	0.05	0.00	0.19	0.07	0.00	0.22	1.4	N/A
	0.06	0.00	0.22	0.08	0.00	0.23	1.33	N/A
RUF	0.07	0.00	0.18	0.12	0.02	0.22	1.71	N/A
	0.09	0.00	0.22	0.13	0.02	0.24	1.44	N/A
AIT	3.36	2.24	3.59	3.40	2.45	3.51	1.01	1.09
	2.95	1.74	3.60	3.18	2.27	3.39	1.07	1.3
SIT	5.22	3.44	5.50	5.14	3.18	5.28	0.98	0.92
	4.09	1.98	5.06	4.72	2.79	4.96	1.15	1.41

Examining weekdays and weekends, correlation trends among traffic features remain similar for either device type.

3) Cross-dimension: Studying correlations across mobility and traffic dimensions, based on subsets of features selected by *CFS*, is a solid step towards an integrated mobility-traffic model. Results are presented in Fig. 9. We find that as the numbers of unique APs/buildings visited (*APC*, *BLD*) *increase*, the average active time (*AAT*), and total and std. of flow counts (*TFC* and *SFC*) *decrease* markedly (significant negative correlation). Surprisingly, there is no noticeable change in total traffic consumed with change in *APC* (negligible correlation), suggesting bundling of more packets in flute flows. (Similar

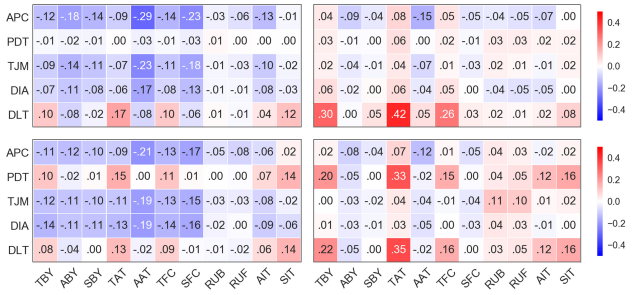


Fig. 9: Correlation plots of mobility vs. traffic on weekdays (top) vs. weekends (bottom) for flutes (left) and cellos (right).

correlation between mobility diameter and the above traffic features) Average IAT (*AIT*) of flutes also rises slightly as mobility metrics *decrease*; for cellos this correlation is almost *nonexistent*. This reinforces our *"stop-to-use"* categorization; cellos are movable but are not active in transit. To sum, *flutes score high on mobility metrics*, have an overall lower flow count and network traffic but produce bigger flows on average. For cellos, on weekends the more time spent at preferred buildings the higher the total active time (*TAT*) and flow counts; this effect exists to a lesser degree for flutes. On weekdays, such correlation does not exist.

B. Steps towards modeling

Here we present our steps towards an integrated mobility-traffic model, with various applications in simulation and protocol design. We utilize daily mobility and traffic features of users during a week[8]. First, we examine how different mobility and traffic features are for flutes and cellos using machine learning. Second, we investigate whether natural convex clusters of users appear in the dataset. These steps verify that the differences of mobility and traffic characteristics across device types are *significant*. We also find that *combining*

[8]Introducing a new detailed model is part of future research.

features from *mobility* ***and*** *traffic* makes this distinction even more *clear*. Finally, mixture models are used to model and synthesize simulated data points of each device type, finding that the accuracy the mixture model *increases* when trained on *combined* features.

1) Supervised classification: Having shown significant differences throughout this study, we used support vector machines (SVM) on different subsets of features to examine the feasibility of device type inference as well as the relationship between mobility and traffic characteristics. These sets include mobility and traffic features *separately*, then *combined*, and then combined with *weekend/weekday labels*. Using *solely mobility features* achieves ≈65% accuracy, while *traffic features alone*, obtains ≈79% accuracy. Using all mobility and traffic variables *combined*, the trained model achieves ≈81% accuracy. Then, as the **combined** feature set is extended to include *weekdays and weekends* independently, the trained SVM yields an accuracy of ≈86%. This suggests that users' behavior (both flutes and cellos) is *more distinguishable* when looking at **combined** mobility and traffic features; especially when *temporal* features such as weekdays are considered separately from weekends. We note that such behavior gaps are *not* the same for both device types and a model should to take that into account.

2) Unsupervised clustering: To investigate natural convex clusters, we used K-means algorithm. Using *mobility features only*, the best mean silhouette coefficient is achieved on k=2 and 4. However, cluster sizes are highly skewed and at k=2, $\approx 60\%$ of devices are correctly clustered. *Traffic features alone*, at k=2, results in $\approx 81.2\%$ accuracy. **Combining** mobility and traffic features, *increases* the accuracy to $\approx 81.5\%$. While some flutes and cellos are similar in terms of mobility and traffic, the clusters of the combined features clearly illustrate **two distinct modes** (especially in *traffic*) and the *high homogeneity* of the clusters hints at *disjoint sets of behaviors* in mobility and traffic dimensions, governed by the device type.

3) Mixture model: To take a step towards synthesis of traces based on our datasets, we trained Gaussian mixture models (GMM) on *combined mobility and traffic features*. From the combined model (CM), we acquired simulated samples. We used Kolmogorov-Smirnov (KS) statistic to compare the simulated samples with the real data and found that CM is able to capture the behaviors of each device type. (Average KS statistic of features is ≈ 0.15 **for flutes** and ≈ 0.14 **for cellos**. See Appendix V in [23] for details.) Importantly, we noted, the combined model produces samples whose *traffic* features match the original data **better**, compared with training a GMM *on traffic features alone* (based on KS statistic), hinting at a key relationship between mobility and traffic. However, comparing mobility features of CM with a GMM trained on mobility features alone shows no improvement to slightly worse results.

Overall, this suggests that there is significant potential for an ***integrated mobility-traffic model*** that captures the differences and ***relationships*** of features, across ***device types***, ***time*** *and* ***space***. We leave detailed comparison of combined modeling with separate modeling of mobility and traffic for future work.

VIII. Conclusion

In this study, we mine large-scale WLAN and NetFlow logs from a campus WiFi network to answer three questions: *(I) How different are mobility and traffic characteristics across device types, time and space? (II) What are the relationships between these characteristics? (III) Should new models be devised to capture these differences? And, if so, how?* We build the *FLAMeS* framework for systematic processing and analysis of the datasets. Using MAC address survey, OUI matching and web domain analysis, we put devices into two categories: flutes (*"on-the-go"*) and cellos (*"stop-to-use"*). We then study a multitude of mobility and traffic metrics, comparing flutes and cellos across time and space. On average, flutes visit twice as many APs as cellos, while cellos generate ≈2x more flows than flutes. However, flutes flows are 2.5x larger in size, with ≈2x the number of packets. The best fit for location preference is **Zipfian**, for flow/packet sizes is **Lognormal**, and for flow IAT at APs is **beta** distribution. Furthermore, flute traffic drops sharply on weekends whereas many cellos remain active. Across mobility and traffic dimensions, we spot a negative correlation for flutes between mobility and flow duration but negligible correlation with traffic size; for cellos, this effect is less pronounced. We find a negative correlation with APs visited and the active time, particularly for flutes. However, no correlation exists APs visited and traffic for cellos. We *quantified* correlations *across both mobility and traffic dimensions*. Finally, we applied machine learning and trained a mixture model to synthesize data points and verified that the **combined** mobility-traffic features capture the *differences* in metrics ***better*** than *either mobility or traffic separately*. Many of our findings are not captured by today's models, and they provide insightful guidelines for the design of evaluation frameworks and simulations models. Hence, this study answered the questions posed, introduced a strong case for newer models, and provided our first step towards a future integrated mobility-traffic model.

IX. Acknowledgements

We would like to thank Prof. Alin Dobra for help in setting up part of the computing cluster, and the anonymous reviewers for their useful feedback. The term *'cello mobility'* was suggested by Prof. Mostafa Ammar and used here with permission. Partial funding for this effort was provided by NSF Award 1320694, August-Wilhelm Sheer fellowship at Technical Univeristy-Munich, and Aalto University.

References

[1] J. Treurniet, "A Taxonomy and Survey of Microscopic Mobility Models from the Mobile Networking Domain," *ACM CSUR*, 2014.

[2] A. Hess, K. A. Hummel, W. N. Gansterer, and G. Haring, "Data-driven Human Mobility Modeling: A Survey and Engineering Guidance for Mobile Networking," *ACM CSUR*, 2016.

[3] D. Kotz and K. Essien, "Analysis of a Campus-Wide Wireless Network," *Springer Wireless Networks*, vol. 11, no. 2, January 2005.

[4] T. Henderson, D. Kotz, and I. Abyzov, "The changing usage of a mature campus-wide wireless network," *Elsevier Computer Networks*, vol. 52, no. 14, October 2008.
[5] G. Maier, F. Schneider, and A. Feldmann, "A First Look at Mobile Hand-held Device Traffic," in *Proc. of IEEE PAM*, 2010.
[6] Y. Zhand and A. Arvidsson, "Understanding the Characteristics of Cellular Data Traffic," in *ACM SIGCOMM CellNet workshop*, 2012.
[7] S. Moghaddam and A. Helmy, "SPIRIT: A simulation paradigm for realistic design of mature mobile societies," in *IWCMC '11*, 2011.
[8] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, 2008.
[9] C. Song, T. Koren, P. Wang, and A.-L. Barabási, "Modelling the scaling properties of human mobility," *Nature Physics*, vol. 6, no. 10, 2010.
[10] D. Zhang, J. Huang, Y. Li, F. Zhang, C. Xu, and T. He, "Exploring human mobility with multi-source data at extremely large metropolitan scales," *MobiCom '14*, 2014.
[11] G. Maier, F. Schneider, and A. Feldmann, "A first look at mobile hand-held device traffic," in *PAM '10*. Springer, 2010.
[12] U. Kumar, J. Kim, and A. Helmy, "Changing patterns of mobile network (WLAN) usage: Smart-phones vs. laptops," *IWCMC '13*, 2013.
[13] X. Chen, R. Jin, K. Suh, B. Wang, and W. Wei, "Network performance of smart mobile handhelds in a university campus wifi network," in *IMC '12*. ACM, 2012.
[14] A. Gember, A. Anand, and A. Akella, "A comparative study of handheld and non-handheld traffic in campus wi-fi networks," in *PAM '11*.
[15] M. Afanasyev, T. Chen, G. M. Voelker, and A. C. Snoeren, "Analysis of a mixed-use urban wifi network: when metropolitan becomes neapolitan," in *SIGCOMM '08*. ACM, 2008.
[16] I. Papapanagiotou, E. M. Nahum, and V. Pappas, "Smartphones vs. laptops: comparing web browsing behavior and the implications for caching," *SIGMETRICS '12*, vol. 40, no. 1, 2012.
[17] H. Falaki, D. Lymberopoulos, R. Mahajan, S. Kandula, and D. Estrin, "A first look at traffic on smartphones," *IMC '10*, 2010.
[18] A. K. Das, P. H. Pathak, C.-N. Chuah, and P. Mohapatra, "Characterization of wireless multidevice users," *ACM Transactions on Internet Technology (TOIT)*, vol. 16, no. 4, pp. 29:1–29:25, Dec. 2016. [Online]. Available: http://doi.acm.org/10.1145/2955096
[19] P. Cao, G. Li, A. Champion, D. Xuan, S. Romig, and W. Zhao, "On human mobility predictability via wlan logs," in *INFOCOM '17*.
[20] A. Nikravesh, Y. Guo, F. Qian, Z. M. Mao, and S. Sen, "An in-depth understanding of multipath TCP on mobile devices," in *MobiCom '16*.
[21] X. G. Meng, S. H. Y. Wong, Y. Yuan, and S. Lu, "Characterizing flows in large wireless data networks," *MobiCom '04*, 2004.
[22] F. Xu, Y. Li, H. Wang, P. Zhang, and D. Jin, "Understanding mobile traffic patterns of large scale cellular towers in urban environment," *IEEE/ACM Transactions on Networking*, vol. 25, no. 2, pp. 1147–1161, April 2017.
[23] "Appendix - Tech. Report." [Online]. Available: https://s3.amazonaws.com/infocom2018/infocom18_flutes_cellos.pdf
[24] A. W. Moore and D. Zuev, "Internet traffic classification using bayesian analysis techniques," *SIGMETRICS '05*, vol. 33, no. 1, 2005.
[25] V. Paxson and S. Floyd, "Wide area traffic: the failure of Poisson modeling," *IEEE/ACM ToN*, vol. 3, no. 3, jun 1995.
[26] T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," *IEEE Communications Surveys & Tutorials*, vol. 10, no. 4, 2008.
[27] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.
[28] W.-J. Hsu, T. Spyropoulos, K. Psounis, and A. Helmy, "Modeling time-variant user mobility in wireless mobile networks."
[29] P. Widhalm, Y. Yang, M. Ulm, S. Athavale, and M. C. González, "Discovering urban activity patterns in cell phone data," *Transportation*, vol. 42, no. 4, pp. 597–623, 2015.
[30] C. Boldrini and A. Passarella, "Hcmm: Modelling spatial and temporal properties of human mobility driven by users' social relationships," *Computer Communications*, vol. 33, no. 9, 2010.

APPENDICES

Here we further describe various aspects of our submitted work to Infocom 2018, which were not included in the original document for brevity.

I. MERGING DATASETS

In order to study network traffic across devices and APs, it is necessary to match the NetFlow records with wireless associations (from WLAN dataset). This task requires the MAC-IP mapping. The IP addresses are dynamically assigned using DHCP but DHCP session logs were not directly available and had to be derived. We define the duration of a DHCP lease as the time between two consecutive associations of a device with any AP; i.e. when a device connects to AP_1, a session starts and once the user device connects to AP_2, the first session ends and a new one starts. Fig. 10 illustrates the associations of a sample device with different APs at different times. The first session would have the IP given by AP_1 and a lease time $t_2 - t_1$, and so on. (total of 5 sessions in this example) The last association is discarded as we do not know the duration of that IP assignment. Combining these derived-DHCP records with the *Location Information* and *Device Type Classification* we create the **DHCP** table.

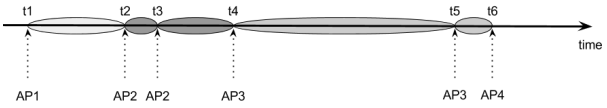


Fig. 10: Wireless association for a device at different times.

The derived DHCP and NetFlow datasets were then merged to form what we refer to as the **CORE** dataset for our study. The unique identifiers between the two are the clients' IPs in addition to start and end time of flows, hence the need for a DHCP-like set. For a DHCP lease session *LS*, all flows whose IP address is the same as the lease *and* whose entire lifetime falls within the lease duration, are associated with *LS*.

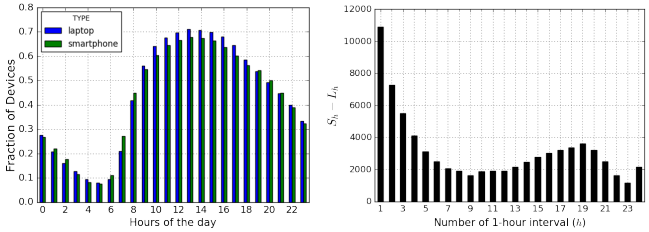
Given these traces, cellular usage cannot be analyzed. However, this does not significantly impact analysis for two reasons: 1) The traces already capture a very large user-base, with tens of thousands of active devices. This raises confidence in our analysis of a real world WLAN. 2) The WiFi campus coverage is ubiquitous, with 1760 APs installed in the vast majority of populated areas. Also, most laptops on campus lack cellular connectivity, and many smartphones use WiFi for their data to avoid cellular data costs.

II. COMPUTING SYSTEM

The size of the datasets is ≈30TB in raw text format, mostly consisting of NetFlow data and ≈0.5TB for AP logs. There were several challenges in managing and mining the largescale datasets that required a thorough preparation, to run on a fast machine with plenty of resources/memory. We explored several techniques and pipelines for extraction, transformation, loading (ETL) and querying of big data and chose tools from Apache Hadoop ecosystem. We use Hive as our data warehouse (tables stored in Parquet format). Apache Spark is the compute engine for data processing and analysis tasks. Computation runs on two nodes, each with 64 cores and ≈0.5TB of memory. Further discussion of the system and comparison to others is out of scope of this document.

III. Mobility Analysis

For completeness, we include further analysis of the mobility aspects of our dataset, discussed in Section V of the conference paper.



(a) (b)

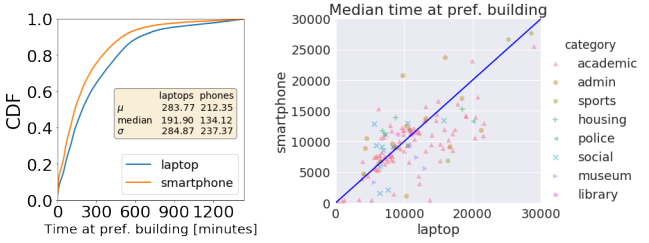
Fig. 11: Hourly associations.

A. Hourly associations

Measuring device associations every hour, Fig. 11a shows the percentage of devices with at least one event as a function of hours of the day. The majority of devices appear online between 9am and 8pm, with the hours between 2am and 6am having less than 20% of devices associating. We find no major differences between flutes' and cellos' distributions, as many users potentially own both. As users arrive on campus and their phones announce their first location, they switch on their laptops. This issue bears further research through a future census study.

To measure the stay of devices throughout a day, we look at 1-hour intervals, and measure the number of hours a device accessed an AP [9]. Fig. 11b depicts $S_h - L_h$, where S_h and L_h are total number of flutes and cellos respectively, with at least one record per hour, as a function of the number of hours online h. Flutes are predominant for short visits and very long stays, but the difference drops significantly at 9 hours, then increases. The rise after 9 hours is likely due to students living on campus, with always-on connected phones.

B. Visitation preferences



(a) (b)

Fig. 12: Time spent at preferred building.

Fig. 12b shows a scatter of the median time spent at a user's preferred building. Each dot represents this value for a given location. This plot shows that *academic*, *police* and *museum* buildings tend to have laptops staying longer, which makes sense intuitively, with students using laptops during lectures and staff working at the other two categories. On the contrary, for *social* and *housing* buildings, there is a higher probability of having flutes staying longer, hinting at a tendency to use mobile devices more in such places. Finally, *administrative*, *sports* and *library* buildings tend to have both types of devices staying for similar amounts of time. Analysis of inherited differences in browsing of online services given by this heterogeneity among buildings is left for future work.

Fig. 12a depicts the time devices spend at their *preferred building* in a day.

C. Return probability

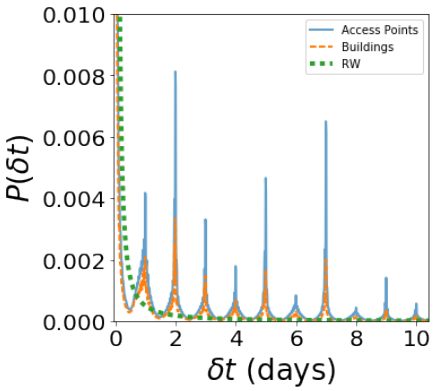


Fig. 13: Probability to return to a previously visited location.

We compare empirical values for devices to return to previously visited APs or buildings in Fig. 13. We observe returning spikes at every 24 hours, with the highest peaks at 48 and 168 hours (2 and 7 days). This can be explained by the schedule of classes at the university.

IV. Traffic Analysis

In this Section, we further discuss references from Section VI of the conference paper.

A. Flow sizes

This metric is the sum of bytes for all packets within a single flow. First, outlier data points are removed using a robust measure of scale, based on inter-quartile range (IQR). Looking at individual flows of each device type shows that size of flows that originated from smartphones are significantly different that laptop flows (p-value< .05). [10]

On weekdays, the average size for smartphone flows is 2070 bytes and 822 bytes for laptop flows; with no significant changes on weekends (CDF in Fig. 14). The difference in medians is more pronounced, on weekdays, for smartphones it is 678 bytes while it is 142 bytes for laptops (similar values on weekends).

[9] P. Widhalm, et al., "Discovering urban activity patterns in cell phone data".

[10] Flow metrics do not fit Gaussian distribution (based on Shapiro-Wilk test for normality, goodness-of-fit test and Q-Q plot results, not included for brevity). Thus, we use Mann-Whitney statistical test[11] to compare two unpaired groups (laptops vs smartphones), and Wilcoxon signed-rank test to compare two paired groups (each device type on weekdays vs weekends) [12].

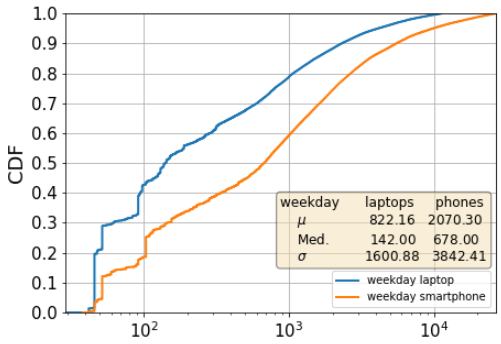


Fig. 14: CDF of individual flow sizes (bytes, log-scale x axis), similar pattern on weekends

B. Lognormal plots

For flow sizes in our dataset, a Lognormal distribution is the best fit, regardless of device type (Fig. 15). Many models assume flow sizes are static, or follow an exponential distribution but real world data provides no supporting evidence. Such simplifying assumptions fail to accurately account for very large flows obtained from a Lognormal distribution.

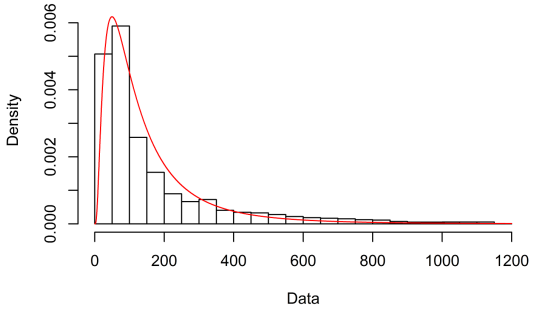


Fig. 15: Lognormal distribution.

C. beta distribution

Inter-arrival times (IAT). Our results show that the flow IAT, regardless of device type, does not follow an

exponential distribution. Flow IAT matches a beta distribution well (Fig. 16) with a very high estimated kurtosis and skewness (estimated at 58 & 6.9 respectively). The high estimated kurtosis illustrates that there are infrequent extreme values, which explains the observed highly elevated standard deviation of IAT[13]

V. First Modeling Steps

More details of the KS test and GMM model are provided here. We found that providing both mobility and traffic features

[13]In the research community, packet IAT and its Fourier transform are considered important features in traffic analysis. They are used extensively in simulation and modeling of networking protocols as well as internet traffic classification [17]. Realistic modeling of IAT is required for accurate simulation and measurement of congestion control mechanisms [24]. Due to limited availability or staleness of most packet-level datasets, although our NetFlow is on a higher abstraction layer (flow-level vs packet-level), analysis of flow IAT can still be used for measuring delay and jitter effects.

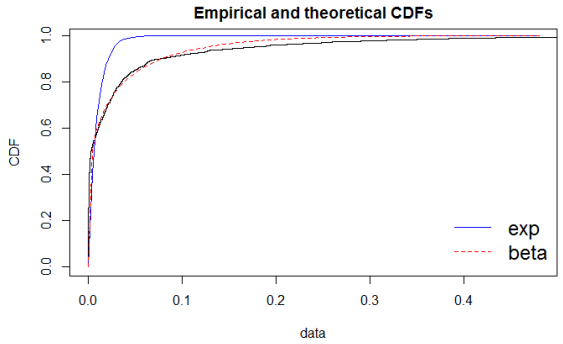


Fig. 16: Exponential and Beta distributions.

TABLE V: KS-statistic summary

KS statistic	Flutes	Cellos
Average	0.150	0.140
Min	0.052	0.027
Max	0.380	0.350
Std	0.086	0.0787

to train a GMM results in lower average KS statistic. Fig 17 shows a sample CDF of *TAT*. KS statistic details can be found in Table V.

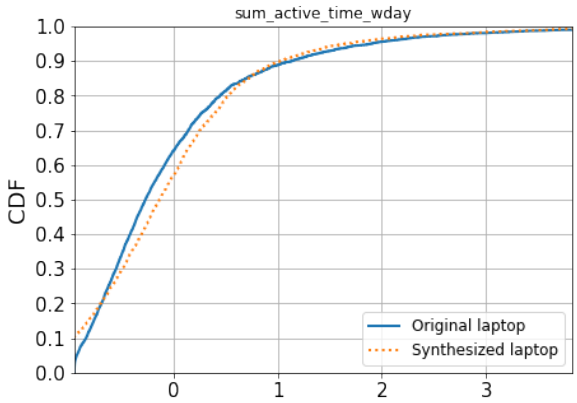


Fig. 17: Synthetic vs. Original *TAT* feature for flutes.

VI. Lessons Learned and Modeling Insights

Our above findings provide further (but surely not yet comprehensive) insights into considerations relevant to the design and parameterization of mobility and network traffic models. While we leave devising and validating a concrete candidate model for future work, we can readily identify the following important elements:

It is crucial to differentiate flutes vs. cellos for both mobility and traffic due to their very different nature. More specifically, flutes exhibit continuous presence whereas cellos are on/off with jumps between locations. Beyond differences in continuity, the traffic patterns (flow sizes, arrival times, etc.) should be specified by device class. Moreover, the traffic generation, spatial locations, and temporal behavior can be linked per device type and per user "community" (e.g. students of different disciplines at various buildings).

Publication 2

©2019 ACM, reprinted with permission from:

Babak Alipour, Leonardo Tonetto, Roozbeh Ketabi, Aaron Yi Ding, Jörg Ott, and Ahmed Helmy. 2019. Where Are You Going Next? A Practical Multi-dimensional Look at Mobility Prediction. In Proceedings of the 22nd International ACM Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWIM '19). Association for Computing Machinery, New York, NY, USA, 5–12. https://doi.org/10.1145/3345768.3355923

ACM Author Gateway

Author Resources

ACM Author Rights

ACM exists to support the needs of the computing community. For over sixty years ACM has developed publications and publication policies to maximize the visibility, impact, and reach of the research it publishes to a global community of researchers, educators, students, and practitioners. ACM has achieved its high impact, high quality, widely-read portfolio of publications with:

- Affordably priced publications
- Liberal Author rights policies
- Wide-spread, perpetual access to ACM publications via a leading-edge technology platform
- Sustainability of the good work of ACM that benefits the profession

Choose

ACM gives authors the opportunity to choose between two levels of rights management for their work. Note that both options obligate ACM to defend the work against improper use by third parties:

- **Exclusive Licensing Agreement:** Authors choosing this option will retain copyright of their work while providing ACM with exclusive publishing rights.
- **Non-exclusive Permission Release:** Authors who wish to retain all rights to their work must choose ACM's author-pays option, which allows for perpetual open access to their work through ACM's digital library. Choosing this option enables authors to display a Creative Commons License on their works.

Post

Otherwise known as "Self-Archiving" or "Posting Rights", all ACM published authors of magazine articles, journal articles, and conference papers retain the right to post the pre-submitted (also known as "pre-prints"), submitted, accepted, and peer-reviewed versions of their work in any and all of the following sites:

- Author's Homepage
- Author's Institutional Repository
- Any Repository legally mandated by the agency or funder funding the research on which the work is based
- Any Non-Commercial Repository or Aggregation that does not duplicate ACM tables of contents. Non-Commercial Repositories are defined as Repositories owned by non-profit organizations that do not charge a fee to access deposited articles and that do not sell advertising or otherwise profit from serving scholarly articles.

For the avoidance of doubt, an example of a site ACM authors may post all versions of their work to, with the exception of the final published "Version of Record", is ArXiv. ACM does request authors, who post to ArXiv or other permitted sites, to also post the published version's Digital Object Identifier (DOI) alongside the pre-published version on these sites, so that easy access may be facilitated to the published "Version of Record" upon publication in the ACM Digital Library.

Examples of sites ACM authors may not post their work to are ResearchGate, Academia.edu, Mendeley, or Sci-Hub, as these sites are all either commercial or in some instances utilize predatory practices that violate copyright, which negatively impacts both ACM and ACM authors.

Distribute

Authors can post an Author-Izer link enabling free downloads of the Definitive Version of the work permanently maintained in the ACM Digital Library.

- On the Author's own Home Page or
- In the Author's Institutional Repository.

Reuse

Authors can reuse any portion of their own work in a new work of their own (and no fee is expected) as long as a citation and DOI pointer to the Version of Record in the ACM Digital Library are included.

- Contributing complete papers to any edited collection of reprints for which the author is notthe editor, requires permission and usually a republication fee.
- Authors can include partial or complete papers of their own (and no fee is expected) in a dissertation as long as citations and DOI pointers to the Versions of Record in the ACM Digital Library are included. Authors can use any portion of their own work in presentations and in the classroom (and no fee is expected).
- Commercially produced course-packs that are sold to students require permission and possibly a fee.

Create

ACM's copyright and publishing license include the right to make Derivative Works or new versions. For example, translations are "Derivative Works." By copyright or license, ACM may have its publications translated. However, ACM Authors continue to hold perpetual rights to revise their own works without seeking permission from ACM.

Minor Revisions and Updates to works already published in the ACM Digital Library are welcomed with the approval of the appropriate Editor-in-Chief or Program Chair.

- If the revision is minor, i.e., less than 25% of new substantive material, then the work should still have ACM's publishing notice, DOI pointer to the Definitive Version, and be labeled a "Minor Revision of"
- If the revision is major, i.e., 25% or more of new substantive material, then ACM considers this a new work in which the author retains full copyright ownership (despite ACM's copyright or license in the original published article) and the author need only cite the work from which this new one is derived.

Retain

Authors retain all perpetual rights laid out in the ACM Author Rights and Publishing Policy, including, but not limited to:

- Sole ownership and control of third-party permissions to use for artistic images intended for exploitation in other contexts
- All patent and moral rights
- Ownership and control of third-party permissions to use of software published by ACM

Copyright © 2024, ACM, Inc

Title: Where Are You Going Next? A Practical Multi-dimensional Look at Mobility Prediction
Authors: Babak Alipour (UFL), **Leonardo Tonetto** (TUM), Roozbeh Ketabi (UFL), Aaron Yi Ding (TU Delft), Jörg Ott (TUM), Ahmed Helmy (UFL)
Venue: 22nd Int'l ACM Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM '19), Miami Beach, FL, USA
Publishing date: November 25–29, 2019
Reference: [3]

Publication Summary

In this article, we focus on the predictability of human mobility patterns using mobile devices. The study aims to investigate the predictability of next location visitation based on different dimensions such as device type, spatial granularity, and temporal spans.

For that, we utilize information-theoretic concepts and machine learning methods to analyze a longitudinal dataset collected from a university campus, which includes fine-grained spatial data and covers a period of 16 months. The two types of devices examined are referred to as "Flutes" (ultra-portable devices like smartphones) and "Cellos" (laptops). The study compares various prediction algorithms and explores the impact of different temporal and spatial granularities on predictability.

We also discuss the concept of entropy estimation using entropy-based estimators (LZ and BWT) as a measure of predictability and describe the different prediction algorithms used in the study, including Markov chains, neural networks (such as LSTM, CNN, and Transformer). We also discuss the dataset used for analysis, which consists of Wi-Fi traces collected from a large university campus.

The experimental results reveal that device type (Flutes vs. Cellos) significantly affects predictability, with Cellos exhibiting higher predictability compared to Flutes across different spatio-temporal granularities. The study also compares the performance of different prediction methods and finds that Cellos are consistently more predictable than Flutes regardless of the algorithm used. We highlight the implications of our findings for predictive caching, user behavior modeling, and mobility simulations.

Overall, the paper aims to contribute to the understanding of human mobility patterns and the development of prediction mechanisms that can be applied in various mobile protocols and applications.

Contribution

As a follow up from the previous paper, I designed the main idea of the paper together with all my co-authors. Using the same data from before, I analyzed the information theory aspects of the paper. I wrote the entire paper, except the parts on the machine learning models, which was done by Babak Alipour. All authors reviewed the text.

Where Are You Going Next? A Practical Multi-dimensional Look at Mobility Prediction

Babak Alipour
University of Florida, USA
babak.ap@ufl.edu

Leonardo Tonetto
Technical University of Munich
Germany
tonetto@in.tum.de

Roozbeh Ketabi
University of Florida, USA
roozbeh@ufl.edu

Aaron Yi Ding
Delft University of Technology
The Netherlands
aaron.ding@tudelft.nl

Jörg Ott
Technical University of Munich
Germany
ott@in.tum.de

Ahmed Helmy
University of Florida, USA
helmy@ufl.edu

ABSTRACT

Understanding and predicting mobility are essential for the design and evaluation of future mobile edge caching and networking. Consequently, research on human mobility prediction has drawn significant attention in the last decade. Employing information-theoretic concepts and machine learning methods, earlier research has shown evidence that human behavior can be highly predictable. Whether high predictability manifests itself for different modes of device usage, across spatial and temporal dimensions is still debatable. Despite existing studies, more investigations are needed to capture intrinsic mobility characteristics constraining predictability, to explore more dimensions (e.g. device types) and spatiotemporal granularities, especially with the change in human behavior and technology.

We investigate practical predictability of next location visitation across three different dimensions: device type, spatial granularity and temporal spans using an extensive longitudinal dataset, with fine spatial granularity (AP level) covering 16 months. The study reveals *device type* as an important factor affecting predictability. Ultra-portable devices such as smartphones have "on-the-go" mode of usage (and hence dubbed "*Flutes*"), whereas laptops are "sit-to-use" (dubbed "*Cellos*"). The goal of this study is to investigate practical prediction mechanisms to quantify predictability as an aspect of human mobility modeling, across time, space and *device types*. We apply our systematic analysis to wireless traces from a large university campus. We compare several algorithms using varying degrees of temporal and spatial granularity for the two modes of devices; *Flutes* vs. *Cellos*.

Through our analysis, we quantify how the mobility of *Flutes* is less predictable than the mobility of *Cellos*. In addition, this pattern is consistent across various spatio-temporal granularities, and for different methods (Markov chains, neural networks/deep learning, entropy-based estimators). This work substantiates the importance of predictability as an essential aspect of human mobility, with direct application in predictive caching, user behavior modeling and mobility simulations.

KEYWORDS

mobility, prediction, markov chain, neural networks, wireless networks, device types

ACM Reference Format:
Babak Alipour, Leonardo Tonetto, Roozbeh Ketabi, Aaron Yi Ding, Jörg Ott, and Ahmed Helmy. 2019. Where Are You Going Next? A Practical Multi-dimensional Look at Mobility Prediction. In *22nd Int'l ACM Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM '19), November 25–29, 2019, Miami Beach, FL, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3345768.3355923

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.
MSWiM '19, November 25–29, 2019, Miami Beach, FL, USA
© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-6904-6/19/11...$15.00
https://doi.org/10.1145/3345768.3355923

1 INTRODUCTION & RELATED WORK

In recent years, large-scale research on human mobility has thrived due to the availability of location data collected from portable computing and communication devices, such as laptops, smartphones, smartwatches and fitness trackers [1]. One particular aspect of human mobility that has gained a lot of attention lately is predictability. Prediction techniques constitute fundamental mechanistic building blocks for many mobile protocols and applications, ranging from resource allocation to caching and recommender systems [2, 3]. In addition, potential improvements to next-hop prediction can lead to more accurate bandwidth predictions, which benefits QoE for users of mobile networks [4].

The seminal work by [5], utilizing cellular network data, established an approach towards understanding and measuring the predictability of human mobility patterns, with their equally important contribution with respect to the data-driven analysis of large mobile populations, and their efforts in devising a framework to study the theoretical limits of predictability. The methods introduced in their framework

are founded in information theory and have since been extensively applied in the area of mobility modeling and prediction. Later studies that built on [5] addressed either the specifics of the prediction problem (e.g., different formulations [6] of the individual's change of location, analyzed different contexts of mobility) or the shortcomings of the original approach (that relied on coarse spatio-temporal granularity). Authors in [7] used Wireless LAN (WLAN) traces from a university campus network and reported multi-modal entropy distributions which can be partially explained by the demographics of the population (*i.e.*, age, gender, major of studies). Other entropy based studies include vehicular mobility [8–10], online social behavior [11, 12], complex systems [13], cellular network traffic [14] and public transport utilization [15]. In addition, the devices' form factor affects the mode of usage and varied traffic profiles ([16–19]), but these studies either do not consider predictability or do not account for different spatio-temporal resolutions. We have chosen our methods based on the literature to measure and compare both theoretical and practical limits of predictability for "on-the-go" *Flutes* and "sit-to-use" *Cellos*, with varying degrees of spatio-temporal granularity, while also looking at the correlation of prediction accuracy with mobility and network traffic profiles using extensive fine-granularity traces (based on our earlier work in [19]).

The *main* questions addressed in this study are: i. How different are *Flutes* and *Cellos* in terms of predictability? ii. How does the predictability of these device types change with different *spatio-temporal granularity* (5, 15, 30 min, 1 hour and 2 hours; access point and building level)? iii. Does the *choice of method* or predictor (*e.g.* Markov Chain, neural networks such as LSTM, CNN and Transformer [20], BWT or LZ based estimators, which are introduced in Section 2) significantly alter the answers to aforementioned questions?

This study provides the following main contributions: 1. Quantifying the differences of *Flutes* and *Cellos* for prediction analysis, evaluated on a real-world large-scale dataset. 2. Comparison of several well-known algorithms (Markov Chains, Neural Networks) and LZ/BWT-based theoretical bounds across different time and space scales for Flutes and Cellos. 3. Use of prediction accuracy as part of the user profile for modeling, and investigation of its correlation with a combination of network traffic and mobility features.

The paper is structured as follows: First, the main approach and methods are presented in Sec. 2. Then, the details of the dataset and experiment setup are discussed in Sec. 3. The experiment results are presented in Sec. 4. Sections 5 and 6 present the discussion on potentials implications of the findings and conclude the paper.

2 MAIN APPROACH & METHODS

We investigate two methods to measure predictability; a theoretical method based on entropy, and a systems method based on practical predictor algorithms. Following we provide the entropy estimation based definition and discuss the different algorithms studied in this paper, including a reference-point Markov Chains approach, and more sophisticated deep learning approaches.

2.1 Entropy Estimation

Entropy is defined as the level of order (or disorder) of a system, and is founded on information theory. It has been adopted in previous studies to establish bounds on predictability under certain assumptions [5, 6]. We utilize it in our study to gauge the performance of our practical predictors. For a random process, this metric is sensitive to both the relative frequency of events and their inter-dependencies [15]. To estimate a baseline of predictability, we compute the *time-uncorrelated* entropy ($\mathcal{S}^{\text{unc}}$) which only takes into account the frequency of the observed events. For the upper-bound of predictability we compute two *time-correlated* estimators based on compression algorithms ($\mathcal{S}^{\text{lz}}$ and $\mathcal{S}^{\text{bwt}}$) which also consider the memory of the system. We define *maximum predictability* as the probability of predicting the most likely state of x_i given a state x_j, which is computed from the entropy S of a given sequence of events based on [5], with the refinements proposed by [6]. For a complete description on *entropy estimation*, we kindly refer the reader to [21, 22].

2.2 Predictors

Markov Chain-based predictor. A Markov chain (MC) with a discrete state space has been applied for user mobility prediction [23, 24]. In an order-k Markov predictor, the state space consists of tuples of k location names (e.g., AP), where the next location prediction depends solely on the most recent preceding k-tuple. We build the model on the data so that observed k-tuples comprise the states. The transition probabilities are learned based on the frequency of appearances of such a transition in observations. The probability for a transition from the current state $S = X_iX_{i+1}...X_j$ to $X_{i+1}X_{i+2}...X_jX_{j+1}$ where $j - i = k$ and each X_i is the symbol for each location, is represented as $PX_{j+1} = c \mid S = X_iX_{i+1}...X_j$ for all c observed in data and is learned based on the reappearance frequency of such a sequence. If the predictor of order k encounters a new sequence that has never seen before, it falls back to the lower, $k-1$ order recursively. The base case is O(0) which is simply the frequency distribution of all symbols observed so far.

Deep learning. Recent approaches to sequence prediction use deep Recurrent Neural Networks (RNN) or Convolutional Neural Networks (CNN). Recurrent neural networks have loops within their cells, allowing information to persist and thus enabling the neural network to connect previous information to make a reasonable prediction of the future state of the modeled system. Certain types of RNNs are capable of learning long-term dependencies. There are multiple variants of RNNs, including Long short-term memory (LSTM) [25] and Gated Recurrent Unit (GRU) [26]. These networks can learn dynamic temporal patterns and have successfully been applied in speech recognition, text-to-speech engines and predicting next location [27, 28].

CNNs learn convolutional filters to extract latent information across the data (i.e. 1D CNNs learn different temporal locality patterns) and use that information for predicting the next location. CNNs have a local receptive field. The receptive field is the region of the input that affects a specific unit of the network, which can be increased by techniques such as stacking more layers.

The Transformer [20] is a novel neural network architecture that only uses self-attention, without any recurrence or convolution, to learn global dependencies between input and output. These networks can be parallelized better (a major shortcoming of RNNs), and also have a global receptive field (as opposed to the local receptive field of CNNs).

In our study, we use a multi-layer LSTM, 1D CNN and a Transformer to predict movements of users based on similar input tuples used for MC-based predictors, as described in the next section. Neural networks are computationally expensive and tend to require hyper-parameter tuning. Thus the deep model is run only on a sample of users in this study. One goal of this study is to analyze the payoff (and cost) of adding complexity to the predictor (e.g. LSTMs), versus the simpler MC-based predictors, while considering different temporal and spatial bins for Flutes vs Cellos.

3 DATASETS & EXPERIMENTAL SETUP

To study the regularity of human behavior, we performed a data-driven analysis applying our methods to a university campus WiFi traces from the University of Florida (UF). The dataset was collected from networks providing wireless access to a large number of portable devices via access points deployed in non-residential areas, including classrooms, computer laboratories, libraries, offices, administrative premises, cafeterias, and restaurants.

Every trace entry contains a unique user identifier (*uuid*), time-stamp and an access point unique identifier (*apid*). Based on the *apid*'s string we are able to identify the building as well as the room in which an access point (AP) was located. Only the geographical coordinates of buildings are known. Table 2 contains a brief summary of the UF dataset with mean (μ) and standard deviation (*std*), where N_{ap} is number of unique access points observed per device, N_{day} number of unique days with at least one record, N_{rec} number of records during data collection, and *total* number of devices available for at least 7 days and accessed more than 5 APs.[1]

3.1 UF traces

The UF traces were collected for 16 months (September/2011-December/2012) and contain over 1700 wireless access points (APs) deployed in 140 buildings which were used by 300K devices. A sample (sythentic) record is shown in Table 1. Its raw records were captured from associations and sessions timeout in which the unique user id (*uuid*) was the MAC address. These *uuid* although hashed, still contained the Organizationally Unique Identifier (OUI)[2] allowing us to distinguish *Flutes* and *Cellos*, as detailed in [19]. This dataset was collected before MAC address randomization became widely available. However, in most current implementations, the randomization only happens in case of probe requests for a network, and once connected to some SSID, the device either presents its original MAC or a generated MAC that does not change per association. Besides, many networks require authentication that allows tracking on higher levels in the network stack (e.g. application). This work is concerned with wireless connectivity being provided to users, and it will always come from discrete points (for example, access points), as opposed to continuous movements in an open field. Thus, all collected WiFi traces are processed as discrete time-series, defined next.

[1] Transient devices are not counted to ensure the analysis is carried out on devices that are mobile and benefit from predictive systems the most, while stationary devices (e.g. plugged-in Cellos) and guests that never return to campus are ignored.

3.2 Discrete-time Series

Given a set a of timely ordered events $X = \{x_t : t = 1, \cdots, n\}$, where x_t is the realization of X at time t for $t \in T$, we say that a timeseries is *discrete* if T are measurements taken at successive times spaced at uniform intervals w, also referred to as sampling rate (defining the temporal granularity).

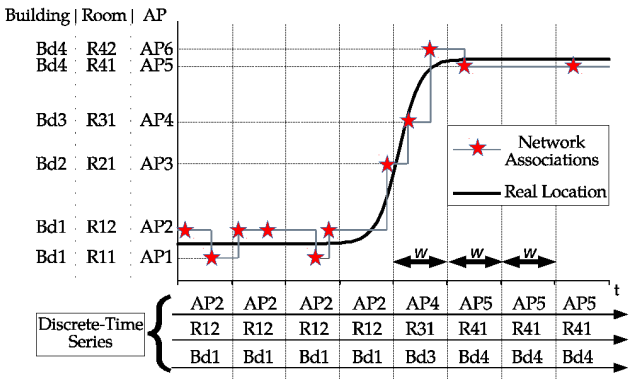


Figure 1: Location of the device is sampled at a constant rate.

Figure 1 depicts an example of how the real location of a device is sensed by the wireless management system through AP associations (red stars) and finally how the discrete-time series is obtained. For a given sampling time window w, our *discrete-time* series may result in different sequences depending on whether we choose an AP or a building as the level of spatial resolution.

From Figure 1, for the first 4 time steps the device switched its associated AP without a real location change. This switch in AP association can be triggered by the mobile device (e.g. stronger wireless signal) or by the network management system (e.g. load balancing).

Note that it is important to define the resolution for *space* and *time*, *i.e.,* how big a location is in space (or point-of-interest) and how often we are going to sample from the input signal. In this example, larger values of w could eliminate this

[2] http://standards.ieee.org/faqs/regauth.html#17

Table 1: AP logs sample data columns

User IP	UUID	AP name	AP MAC	Lease begin time	Lease end time
10.130.90.3	00:11:22:00:00:00	b422r143-win-1	00:1d:e5:8f:1b:30	1333238737	1333238741

Table 2: Statistics per device available for at least 7 days & accessed more than 5 APs.

	N_{ap}		N_{day}		N_{rec}		Total Devices
	μ	std	μ	std	μ	std	
UF	127.3	142.3	63.5	59.2	1861	5121	138028

ping-pong effect of switching between APs without actually moving, but also cause loss of information when the user transits from one location to another. On the contrary, very small values of w could over-sample long periods when the user is not moving. Similarly, different values of spatial resolution could mitigate noise but eliminate information from the traces. Choosing these parameters is often influenced by the characteristics of the available dataset as well as the targeted application of the study.

Step Value. A weighing mechanism is used to pick the corresponding location to represent a time step. During a time interval, we weigh every observed location of the device with the duration of time at that location and pick the one with the highest weight to represent that step. We assign a user to a specific location ℓ in the time interval δt between an association at ℓ and the next association at any other location, but only if $\delta t < t_{\max}$. After $t_{\max}$ the device will be in an *unknown* state [5] until the next network event which will reveal its location for future steps.

3.3 Experiments

The design of our experiments is based on our study's questions: i. How different are *Flutes* and *Cellos* in terms of predictability? ii. How does the predictability of these device types change with different spatio-temporal granularity? iii. Does the choice of method or predictor significantly alter the answers to the aforementioned questions? Thus, we evaluated a matrix, involving *combinations* of the following dimensions:

- Device Types: *Flutes* vs. *Cellos*.
- Temporal Resolutions: 5 min, 15 min, 30 min, 1 hour and 2 hours.
- Spatial Resolutions: Access Points, and Buildings.
- Methods: A. Well-known sequence prediction algorithms from machine learning literature (Markov Chains, Neural Networks) B. Entropy-based Estimations of predictability upper-bounds.

The temporal resolutions are chosen based on the related literature, and the spatial resolutions are determined by the granularity of the dataset. The experiments were implemented in Python, the neural networks were implemented using Tensorflow [3] and Keras. Training is carried out in an *online* manner and the evaluation is through providing a sliding window of k observations to the predictor and testing the prediction correctness of the next symbol. The *fraction of correct next symbol predictions*, or success rate, is the prediction accuracy metric.

[3]TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org.

4 EXPERIMENTAL RESULTS

4.1 Spatio-Temporal Resolutions

To answer the first two questions of this study, particularly "ii. How does the predictability of these device types change with different *spatio-temporal granularity*?", Table 3 summarizes the median accuracy of an LSTM predictor for Flutes and Cellos with different spatial and temporal granularity.

The choice of granularity is application-dependent, for example, to predict foot traffic at buildings and congestion planning based on density, building level analysis is more appropriate. Cellos show more predictable behavior overall, as the fraction of correct next symbol predictions is higher for Cellos across the board. At the AP level, with longer time bins, the accuracy for both Flutes and Cellos decreases. This observation is in line with previous findings [6]. At 15min time intervals, the difference between Flutes and Cellos is at its maximum, then drops and remains stable for longer time intervals. At the building level, the accuracy follows a less regular pattern but both Flutes and Cellos are most predictable at 5min intervals (mainly due to long repeats of the same location in the sequence). Cellos' accuracy drops for 30min bins and goes back up again. On the other hand, Flutes are more predictable in 30min bins than 15min, 1h or 2h bins.

Looking across all temporal bins, Fig 2 presents the empirical cumulative distribution function (ECDF) of prediction accuracy at AP and building spatial granularity. The "sit-to-use" *Cellos* show significantly higher predictability at every percentile; this is reasonable given their lower mobility [19] and mode of usage. In fact, prediction accuracy is highly correlated with other mobility and network traffic features of mobile wireless users, we will take a brief look at these correlations in Section 5 and Fig 4.

4.2 Comparison of Methods

To answer the third question of this study, "iii. Does the choice of method or predictor significantly alter the answers to the aforementioned questions?", here we compare the experiment results for different methods: 1) *MC*: Markov Chain 2) *LSTM*: A type of recurrent neural network 3) *CNN*: 1D convolutional neural network 4) *Transformer*: A type of self-attention neural network 4) *Hr_LZ*: Theoretical predictability based on the Lempel-Ziv (LZ) entropy estimator 5) *Hr_BWT*: Theoretical predictability based on the Burrows-Wheeler transform (BWT) entropy estimator. A summary of

Table 3: Median accuracy percentages of LSTM (sequence len. 40) for *Flutes* vs *Cellos*, 5min-2h temporal and AP/Bldg spatial granularity.

	AP		Building	
	F	C	F	C
5 min	33.22	42.25	44	63.4
15 min	21.42	36.9	34.53	58.06
30 min	21.88	27.39	39.56	50.78
1 hour	19.67	24.33	32.62	52.03
2 hour	17.17	22.5	32.6	59.62

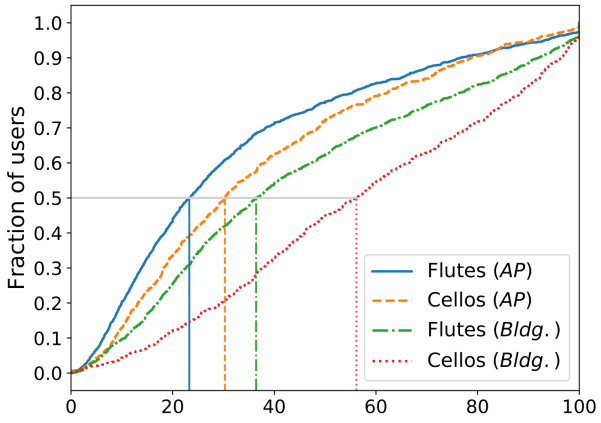


Figure 2: ECDF of LSTM Prediction Accuracy for *Flutes* & *Cellos* at AP and Building spatial levels (all temporal levels combined, vertical lines denote medians, sequence length 40).

comparisons is presented in Table 4, for temporal granularity of 1h and 15min, highlighting the difference of *Cellos* - *Flutes*.

In all cases *Cellos* are more predictable than *Flutes*, regardless of the choice of method (with a minor exception of LZ predictor at 15 minutes time and building level which might be due to intrinsic instability of LZ based estimator). The difference in median accuracy for *Flutes* vs *Cellos* is up to 25% (Building level, 15 minutes window, sequence length 40, *Flutes* 33.97% vs *Cellos* 59.03%). Other temporal choices result in a similar pattern. Another notable observation is that while the neural networks are more complex, and require vastly more computing power, they only achieve modest increase compared to Markov Chains in *some* scenarios (e.g., *Cellos*, at the building level and sequence length 40, from 48.56% to 52.5%). This is a trade-off that needs to be considered in the design of predictive caching systems. In addition, increasing the sequence length k (i.e. the number of previous time steps available to the predictor) impacts the Markov Chain model more than the neural networks. This is particularly pronounced for 15 minutes time window, in fact, the neural networks do not lose much accuracy from increasing sequence length 5 to 40 in case of the 1 hour time window. Also, the theoretical LZ and BWT based estimators, show higher upper bounds compared with the best of the algorithms, with sequence length 5 Markov Chains and CNNs being the closest practical algorithms for the 15 minutes case. The predictors are far behind in the 1h case, suggesting room for improvement via tuning for specific time and space granularities. The run time of LSTM is the longest, followed by CNN (not shown for brevity). In addition, in case of the Transformer, at 1 hour temporal resolution, median accuracy is slightly higher compared to LSTM in most cases. However, in the shorter 15 minute resolution, the accuracy is significantly better for Flutes (average accuracy ≈14% higher than LSTM), and slightly better for Cellos. This shows the utility of adapting advances in deep learning to mobility prediction.

4.3 Top 2 Locations

In order to improve the obtained success rate in predicting the next location, we evaluated our prediction methods when considering the top 2 possible locations. In other words, we evaluate the accuracy of the predictors when considering not only the best possible location but the two places where the user is most likely to be found in the next time slot. In this case, we are interested in assessing this improvement which could be beneficial for preemptive caching systems.

Overall, we observe an increase of up to 20% in the median accuracy of all predictors evaluated. Figure 3 depicts the differences between the top 1 and top 2 for CNN's in different temporal and spatial levels. Interestingly, more pronounced improvements were observed at higher spatial levels (buildings) where top 1 accuracy was already higher. The upward trend continues when measuring the top 3 accuracy, though it is less dramatic. The change in accuracy, of top 1 to top 3, for LSTMs followed a similar pattern.

These improvements could be explained by the expected uncertainty in choosing where to go next being better described by more than one location. When deciding between these multiple options, a user is likely to use information not available in our mobility traces. Therefore, when asking our predictor for the next step with the highest probability, these top locations would seem random, and allowing even a small number of top choices (> 1) greatly improves its success rate.

To numerically support this conjecture, we look into the average uncertainty in picking a next location given by $U_{\text{next}} = 2^{S_{\text{rate}}}$, where S_{rate} is the entropy rate estimated, for which we used the BWT algorithm (S^{bwt}, see Section 2.1). For a user's sequence of visited locations, this metric summarizes the average uncertainty about the user's next step at every location, therefore the higher this number the more random the next steps seem to be for a given pair of spatial and temporal levels. Table 5 presents the expected U_{next}, for both Flutes (**F**) and Cellos (**C**). Interestingly, these values not only correlate with the obtained values for accuracy but also shows a clear correspondence with the increase in accuracy when using the top 2. For example, at the AP and 1-hour levels we observe a high U_{next} as well as a marginal improvement from top 1 to top 2, while in contrast at building and 15 minutes levels U_{next} are lower and the improvements for our predictor accuracy are more pronounced.

Table 4: Summary of Median Accuracy for *Flutes* vs *Cellos* with different methods (Diff is $Cellos - Flutes$) and sequence lengths for 15min and 1h time windows.

		AP, 1h			Bldg., 1h			AP, 15min			Bldg., 15min		
Seq Len	Predictor	F	C	Diff	F	C	Diff	F	C	Diff	F	C	Diff
5	MC	21.05	25.95	+4.90	38.25	53.50	+15.25	61.72	70.30	+8.58	75.00	87.60	+12.60
	LSTM	21.62	25.00	+3.38	35.03	50.00	+14.97	40.00	44.56	+4.56	52.44	65.56	+13.12
	CNN	16.45	24.27	+7.82	34.94	50.00	+15.06	50.00	59.80	+9.80	64.60	76.94	+12.34
10	MC	17.98	25.6	+7.62	36.72	50.28	+13.56	52.25	61.97	+9.72	68.00	82.25	+14.25
	LSTM	20.83	26.31	+5.48	37.50	50.66	+13.16	31.14	44.62	+13.48	45.38	64.56	+19.18
	CNN	18.06	22.62	+4.56	36.20	52.03	+15.83	49.20	58.80	+9.60	64.56	74.00	+9.44
20	MC	18.1	24.52	+6.42	36.28	49.94	+13.66	38.50	48.22	+9.72	57.30	74.94	+17.64
	LSTM	21.22	24.19	+2.97	36.12	50.78	+14.66	29.17	41.00	+11.83	43.62	61.47	+17.85
	CNN	18.44	23.60	+5.16	35.28	50.00	+14.72	37.84	48.12	+10.28	50.00	65.00	+15.00
40	MC	17.88	23.61	+5.73	35.1	48.56	+13.46	27.97	31,00	+3.03	47.12	65.80	+18.68
	LSTM	19.67	24.33	+4.66	32.62	52.03	+19.41	23.30	39.40	+16.10	33.97	59.03	+25.06
	CNN	18.75	23.97	+5.22	35.25	52.50	+17.25	27.62	44.70	+17.08	41.25	62.10	+20.85
	LZ	46.90	52.60	+5.70	58.78	66.40	+7.62	72.70	76.06	+3.36	79.60	79.10	-0.50
	BWT	66.44	69.44	+3.00	73.70	79.90	+6.20	83.30	88.06	+4.76	88.60	92.20	+3.60

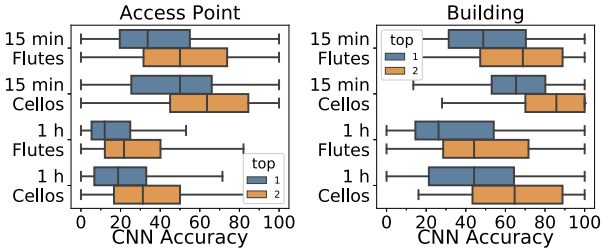


Figure 3: CNN accuracy for top 1 and top 2 locations.

Table 5: User's expected uncertainty μ when choosing next location ($U_{\text{next}} = 2^{S_{\text{rate}}}$). Error given by standard deviation σ.

		AP		Building	
		$\mu \pm \sigma$	*95th-%*	$\mu \pm \sigma$	*95th-%*
15 minutes	**F**	3.10 ± 1.3	5.3	2.17 ± 0.7	3.3
	C	2.05 ± 0.7	3.3	1.56 ± 0.4	2.2
1 hour	**F**	5.50 ± 2.4	9.7	3.65 ± 1.7	6.5
	C	3.48 ± 1.6	6.37	2.10 ± 0.9	3.7

These findings show one of the trade-offs a predictive caching system would need to consider, that is to find the balance between the number of places to prefetch assets and the desired level of cache hit ratio.

5 DISCUSSION & FUTURE WORK

In this paper, we define our research problem as predicting the next symbol in a discrete-time series for users with two categories of devices. The next symbol either denotes the next access point or building in the visitation sequence. The accuracy is evaluated as the fraction of the next symbols predicted correctly.

While some earlier studies investigated a similar problem setup, our study has notable implications. For example, across device types, predictability can vary significantly, with Cellos showing typically higher predictability. Also, with larger time windows such as 1 hour, it is easy to miss short stays (since one location visit with a duration of 31 minutes would result in other locations in that 1 hour window being ignored). On the other hand, a short time window results in multiple repetitions of the same location in the sequence, potentially achieving high prediction accuracy even when the method is not predicting the *transitions* well. Further, we also note that allowing prediction algorithms to look further back does not help prediction in most cases; this might be an artifact of the users' likelihood to stay in place over limited time spans, which makes predicting a 'stay' straightforward while predicting a location transition remains challenging.

Our results highlight the importance of considering the device type, context, and application in order to choose an appropriate time and space granularity; the best performing method differs across these dimensions. Furthermore, we observe a significant increase in accuracy, of up to 20%, when considering the top 2 possible next locations compared to only measuring top 1 accuracy, highlighting the complexity of these predictions based only in the history of visits from

a user. In some cases, such as 1 hour, access point level prediction, the median of the top 2 accuracy of the population is nearly twice as high as the median top 1 accuracy. Many misclassifications occur because the prediction algorithm is simply confused between only two places. For certain applications, such as predictive caching, it can be worthwhile to consider preloading in more than one location to improve the user experience at the expense of increased resource consumption, a trade-off to measure in future studies.

Interesting possible problems yet to be addressed include, taking the distance between possible locations into account when selecting a future stop, as well as cluster users with similar mobility patterns to further improve the prediction accuracy of their movements.

All the findings here are based on the university dataset (Sec. 3), which provides a peek into only a subset of the population, so we emphasize the importance of reproducing these analyses on other datasets in different settings.

Correlations with Mobility and Network Traffic. Figure 4 shows the correlation of prediction accuracy with a sample of features that describe the mobility or network traffic of users. PDT(W/E) and TJ(W/E) are mobility features while AAT(W/E) and AI(W/E) are traffic features. PDTW is the time spent at the user's preferred building (most common) on weekdays (PDTE for weekends). TJW is the total sum of jumps (distance) for the weekdays while TJE describes the same feature for weekends. AATW is the average of active time (as indicated by network usage) of the user for weekdays (AATE for weekends). AIW stands for the average inter-arrival time of flows on weekdays, and AIE for weekends ([19, 29]).

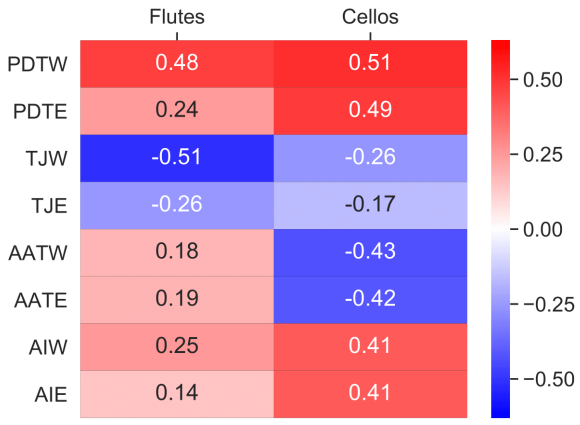


Figure 4: Pearson Correlation of Prediction Accuracy with several Mobility and Network Traffic Features.

The results present significant correlations between the prediction accuracy, with not only the *mobility features*, but also *network traffic features*. These correlations vary across device types (*Flutes* vs *Cellos*), and in time (*Weekdays* vs *Weekends*). This is a very important observation for the design of *predictive caching* systems, importantly, it might be possible to improve prediction of *where* the user is going based on network traffic profile while noting the different modes of usage based on device types. We leave the investigation of incorporating this extra information and potential improvements to future work.

Integrated Mobility-Traffic Modeling. Given the observed correlations, we hypothesize that the use of *predictability* as a feature in an integrated mobility-traffic generative model could lead to more realistic synthetic traces. Such a data-driven generative model would be an essential tool for network simulations and capacity planning. Notably, it can also be made *privacy preserving*, since collected traces would be replaced with realistic synthetic data that captures mobility, network traffic, predictability, and their relationships. Further study is beyond the scope of this work and is left for future work.

6 CONCLUSION

In this work, we sought to answer three questions: i. How different are *Flutes* and *Cellos* in terms of predictability? ii. How does the predictability of these device types change with different *spatiotemporal granularity*? iii. Does the *choice of method* or predictor significantly alter the answers to the aforementioned questions? For this purpose, we processed a large-scale dataset from a campus environment, and grouped the devices into two categories; and chose a set of methods for the comparisons including Entropy-based estimators and popular algorithms such as Markov Chains and Neural Networks.

The results of experiments show the movements of *Cellos* ("sit-to-use") are significantly more predictable than *Flutes* (up to 25% difference in accuracy). This pattern is consistent across various temporal granularities (5 min to 2 hours), spatial granularities (Access Point and Building level), and for different methods (Markov Chains, Neural Networks, Entropy-based Estimators). We illustrate that the performance of predictors depends strongly on the span of temporal bins. Markov Chains tend to outperform deep learning models in shorter time-bins while LSTMs and CNNs usually show a higher accuracy in longer time-bins. CNNs have mostly similar accuracy to LSTMs in the latter case but have significantly better run time on a modern GPU. Furthermore, looking at the top 2 locations we observe an increase of up to 20% suggesting that higher accuracy is achievable when considering multiple possible next locations.

We also found significant correlations among prediction accuracy, *mobility features*, and also *network traffic features*, varying across device types, an important observation for the design of predictive caching systems where it might be possible to improve mobility prediction based on network traffic profile. We plan to further investigate the use of *predictability as a feature* in an integrated mobility-traffic generative model, and its application in state-of-the-art predictive caching systems.

ACKNOWLEDGEMENT

This work was partially funded by NSF Award 1320694. We gratefully acknowledge the support of NVIDIA Corp. with the donation of the Titan Xp GPU used for this research.

REFERENCES

[1] K. Jayarajah, R. K. Balan, M. Radhakrishnan, A. Misra, and Y. Lee, "Livelabs: Building in-situ mobile sensing & behavioural experimentation testbeds," in *MobiSys*. ACM, 2016.

[2] V. Siris, X. Vasilakos, and D. Dimopoulos, "Exploiting mobility prediction for mobility, popularity caching and dash adaptation," in *WoWMoM*, 2016.

[3] N. Lathia, "The anatomy of mobile location-based recommender systems," in *Recommender Systems Handbook*. Springer, 2015.

[4] T. Mangla, N. Theera-Ampornpunt, M. Ammar, E. Zegura, and S. Bagchi, "Video through a crystal ball: Effect of bandwidth prediction quality on adaptive streaming in mobile environments," in *MoVid*. ACM, 2016.

[5] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, 2010.

[6] G. Smith, R. Wieser, J. Goulding, and D. Barrack, "A refined limit on the predictability of human mobility," *PerCom*, 2014.

[7] P. Cao, G. Li, A. Champion, D. Xuan, S. Romig, and W. Zhao, "On human mobility predictability via WLAN logs," in *Proc. INFOCOM*, Apr. 2017.

[8] Y. Li, D. Jin, P. Hui, Z. Wang, and S. Chen, "Limits of predictability for large-scale urban vehicular mobility," *IEEE T-ITS*, 2014.

[9] J. Wang, Y. Mao, J. Li, Z. Xiong, and W. X. Wang, "Predictability of road traffic and congestion in urban areas," *PLoS ONE*, 2015.

[10] R. Gallotti, A. Bazzani, M. D. Esposti, and S. Rambaldi, "Entropic measures of individual mobility patterns," *JSTAT*, 2013.

[11] T. Takaguchi, M. Nakamura, N. Sato, K. Yano, and N. Masuda, "Predictability of conversation partners," *Physical Review X*, 2011.

[12] R. Sinatra and M. Szell, "Entropy and the predictability of online life," *Entropy*, vol. 16, no. 1, pp. 543–556, 2014.

[13] R. Hanel and S. Thurner, "A comprehensive classification of complex statistical systems and an axiomatic derivation of their entropy and distribution functions," *Epl*, vol. 93, no. 2, 2011.

[14] X. Zhou, Z. Zhao, R. Li, Y. Zhou, and H. Zhang, "The predictability of cellular networks traffic," in *ISCIT 2012*, 2012.

[15] G. Goulet-Langlois, H. N. Koutsopoulos, Z. Zhao, and J. Zhao, "Measuring regularity of individual travel patterns," *IEEE T-ITS*, 2017.

[16] G. Maier, F. Schneider, and A. Feldmann, "A first look at mobile hand-held device traffic," in *PAM*. Springer, 2010.

[17] X. Chen, R. Jin, K. Suh, B. Wang, and W. Wei, "Network performance of smart mobile handhelds in a university campus wifi network," *ACM IMC*, 2012.

[18] U. Kumar, J. Kim, and A. Helmy, "Changing patterns of mobile network (WLAN) usage: Smart-phones vs. laptops," *IWCMC*, 2013.

[19] B. Alipour, L. Tonetto, A. Yi Ding, R. Ketabi, J. Ott, and A. Helmy, "Flutes vs. cellos: Analyzing mobility-traffic correlations in large wlan traces," in *IEEE INFOCOM*, 2018.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[21] H. Cai, S. R. Kulkarni, and S. Verdú, "Universal entropy estimation via block sorting," pp. 1551–1561, 2004.

[22] Y. Gao, I. Kontoyiannis, and E. Bienenstock, "Estimating the entropy of binary time series: Methodology, some theory and a simulation study," *Entropy*, vol. 10, no. 2, pp. 71–99, 2008.

[23] L. Song, D. Kotz, R. Jain, and X. He, "Evaluating location predictors with extensive Wi-Fi mobility data," in *INFOCOM*, 2004.

[24] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the limit of predictability in human mobility," *Scientific reports*, vol. 3, 2013.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv:1406.1078*, 2014.

[27] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[28] A. Karatzoglou, A. Jablonski, and M. Beigl, "A seq2seq learning approach for modeling semantic trajectories and predicting the next location," in *ACM SIGSPATIAL*, 2018.

[29] B. Alipour, M. Al Qathrady, and A. Helmy, "Learning the relation between mobile encounters and web traffic patterns: A data-driven study," in *ACM MSWIM*, 2018.

Publication 3

©2019 ACM, reprinted with permission from:

Leonardo Tonetto, Moritz Untersperger, and Jörg Ott. 2019. Towards Exploiting Wi-Fi Signals From Low Density Infrastructure for Crowd Estimation. In Proceedings of the 14th Workshop on Challenged Networks (CHANTS'19). Association for Computing Machinery, New York, NY, USA, 27–32. https://doi.org/10.1145/3349625.3355439

ACM Author Gateway

Author Resources

ACM Author Rights

ACM exists to support the needs of the computing community. For over sixty years ACM has developed publications and publication policies to maximize the visibility, impact, and reach of the research it publishes to a global community of researchers, educators, students, and practitioners. ACM has achieved its high impact, high quality, widely-read portfolio of publications with:

- Affordably priced publications
- Liberal Author rights policies
- Wide-spread, perpetual access to ACM publications via a leading-edge technology platform
- Sustainability of the good work of ACM that benefits the profession

Choose

ACM gives authors the opportunity to choose between two levels of rights management for their work. Note that both options obligate ACM to defend the work against improper use by third parties:

- **Exclusive Licensing Agreement:** Authors choosing this option will retain copyright of their work while providing ACM with exclusive publishing rights.
- **Non-exclusive Permission Release:** Authors who wish to retain all rights to their work must choose ACM's author-pays option, which allows for perpetual open access to their work through ACM's digital library. Choosing this option enables authors to display a Creative Commons License on their works.

Post

Otherwise known as "Self-Archiving" or "Posting Rights", all ACM published authors of magazine articles, journal articles, and conference papers retain the right to post the pre-submitted (also known as "pre-prints"), submitted, accepted, and peer-reviewed versions of their work in any and all of the following sites:

- Author's Homepage
- Author's Institutional Repository
- Any Repository legally mandated by the agency or funder funding the research on which the work is based
- Any Non-Commercial Repository or Aggregation that does not duplicate ACM tables of contents. Non-Commercial Repositories are defined as Repositories owned by non-profit organizations that do not charge a fee to access deposited articles and that do not sell advertising or otherwise profit from serving scholarly articles.

For the avoidance of doubt, an example of a site ACM authors may post all versions of their work to, with the exception of the final published "Version of Record", is ArXiv. ACM does request authors, who post to ArXiv or other permitted sites, to also post the published version's Digital Object Identifier (DOI) alongside the pre-published version on these sites, so that easy access may be facilitated to the published "Version of Record" upon publication in the ACM Digital Library.

Examples of sites ACM authors may not post their work to are ResearchGate, Academia.edu, Mendeley, or Sci-Hub, as these sites are all either commercial or in some instances utilize predatory practices that violate copyright, which negatively impacts both ACM and ACM authors.

Distribute

Authors can post an Author-Izer link enabling free downloads of the Definitive Version of the work permanently maintained in the ACM Digital Library.

- On the Author's own Home Page or
- In the Author's Institutional Repository.

Reuse

Authors can reuse any portion of their own work in a new work of their own (and no fee is expected) as long as a citation and DOI pointer to the Version of Record in the ACM Digital Library are included.

- Contributing complete papers to any edited collection of reprints for which the author is notthe editor, requires permission and usually a republication fee.
- Authors can include partial or complete papers of their own (and no fee is expected) in a dissertation as long as citations and DOI pointers to the Versions of Record in the ACM Digital Library are included. Authors can use any portion of their own work in presentations and in the classroom (and no fee is expected).
- Commercially produced course-packs that are sold to students require permission and possibly a fee.

Create

ACM's copyright and publishing license include the right to make Derivative Works or new versions. For example, translations are "Derivative Works." By copyright or license, ACM may have its publications translated. However, ACM Authors continue to hold perpetual rights to revise their own works without seeking permission from ACM.

Minor Revisions and Updates to works already published in the ACM Digital Library are welcomed with the approval of the appropriate Editor-in-Chief or Program Chair.

- If the revision is minor, i.e., less than 25% of new substantive material, then the work should still have ACM's publishing notice, DOI pointer to the Definitive Version, and be labeled a "Minor Revision of"
- If the revision is major, i.e., 25% or more of new substantive material, then ACM considers this a new work in which the author retains full copyright ownership (despite ACM's copyright or license in the original published article) and the author need only cite the work from which this new one is derived.

Retain

Authors retain all perpetual rights laid out in the ACM Author Rights and Publishing Policy, including, but not limited to:

- Sole ownership and control of third-party permissions to use for artistic images intended for exploitation in other contexts
- All patent and moral rights
- Ownership and control of third-party permissions to use of software published by ACM

Copyright © 2024, ACM, Inc

Title: Towards Exploiting Wi-Fi Signals From Low Density Infrastructure for Crowd Estimation
Authors: **Leonardo Tonetto** (TUM), Moritz Untersperger (TUM), Jörg Ott (TUM)
Venue: 14th Workshop on Challenged Networks (CHANTS'19), Los Cabos, Mexico
Publishing date: October 25, 2019
Reference: [4]

Publication Summary

In this paper, we explore the use of Wi-Fi signals for crowd estimation while preserving privacy. We highlight the importance of crowd assessment in various applications but note that traditional methods, such as cameras, can compromise privacy. To address this issue, we propose using the received signal strength (RSS) of Wi-Fi signals from stationary beacons to estimate crowd density.

We discuss the use of wireless devices like smartphones, tablets, and laptops, equipped with sensors, to gather information about crowd behavior. However, the collection of such behavioral data without user consent raises privacy concerns. To ensure privacy, we propose using Wi-Fi signals and management frames to estimate crowd size without tracking individuals.

We present our approach for estimating crowd size based on RSS values. We discuss signal propagation models, such as the free-space path loss (FSPL) model and the log-distance path loss (LDPL) model, to understand the signal attenuation caused by human presence. We propose a fixed sender model, where the path loss between a fixed sender and a fixed measuring probe in an empty building is compared to the path loss with a crowd present. The difference in path loss is assumed to be proportional to the number of people in the monitored area.

We also explain the use of IEEE 802.11 Management Frames in crowd estimation where MAC address randomization techniques may affect device counting accuracy and that the variable number of devices per person can introduce inaccuracies in crowd size estimation.

In the evaluation section, we apply our RSS-based method to a real-world measurement in a large building. We observe a strong linear relationship between the average path loss and the number of mobile devices counted. This finding suggests that crowd size can be estimated based on the RSS values from stationary devices without compromising privacy.

Overall, the we present a privacy-preserving approach to crowd estimation using Wi-Fi signals. By utilizing RSS values from stationary beacons, we demonstrate the feasibility of estimating crowd density without tracking individuals. The proposed method provides a valuable contribution to the field of crowd assessment and has applications in various domains such as disaster management, network evaluation, and human mobility modeling.

Contribution

I came up with the idea with supervision from Jörg Ott, designed the experiments and wrote the manuscript. Moritz collected and analyzed the data. All authors reviewed the text.

Towards Exploiting Wi-Fi Signals From Low Density Infrastructure for Crowd Estimation

Leonardo Tonetto
tonetto@in.tum.de
Technical University of Munich
Germany

Moritz Untersperger
moritz.untersperger@tum.de
Technical University of Munich
Germany

Jörg Ott
ott@in.tum.de
Technical University of Munich
Germany

ABSTRACT

The ubiquity of wireless devices such as smartphones, tablets and laptops, has enabled sensing large crowds. This was made possible with numerous methods available that mostly listen to Bluetooth or Wi-Fi channels to observe traffic diversity, sources, and destinations. On one hand, it is clearly useful to create crowd awareness, for example to estimate the number of people and assess people flows inside buildings or in areas, with applications in disaster management, network evaluation, and human mobility modeling, as well as for individual mobile devices to assess their context. At the same time, most of these network activity monitoring methods risk compromising the privacy of the individuals being counted and possibly—deliberately or inadvertently—tracked. That is, they may leak private information about people's individual mobility patterns without their consent or even awareness. In this paper, we take a first stab at addressing the problem of privacy-preserving crowd (density) estimation by utilizing the received signal strength (RSS) of Wi-Fi signals from stationary beacons. We use management frames as an approximation of ground truth to validate our observations. We evaluate this method in a real world measurement, observing very strong correlations between the presence of over 35,000 mobile devices in a large building and Wi-Fi RSS values from stationary devices.

KEYWORDS

crowd assessment, privacy, wireless networks

ACM Reference Format:
Leonardo Tonetto, Moritz Untersperger, and Jörg Ott. 2019. Towards Exploiting Wi-Fi Signals From Low Density Infrastructure for Crowd Estimation. In *CHANTS '19: The ACM Workshop on Challenged Networks October 21–25, 2019, Los Cabos, Mexico.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

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 ACM 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.
CHANTS '19, October 21–25, 2019, Los Cabos, Mexico
© 2019 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Crowd assessment is of great importance for managing large events, human mobility modeling for infrastructure deployment as well as for feeding simulations such as pedestrians in urban environments or, simply, to be aware of our surroundings. Common commercial solutions use cameras which can compromise the privacy of the tracked subjects.

Mobile wireless devices, such as smartphones, tablets and laptops have enabled detailed research about crowd behavior in the recent years. Equipped with various sensors, these devices provide researchers rich information about determined actions or states of its users.

This behavioral information however, may be sensed without the users' consent. To prevent this, many countries have changed their laws to strictly regulate how such sensitive data (such as users' location) should be handled. Therefore, a crowd assessment system is needed which is capable of preserving the privacy of its subjects while accurately estimating the size of a group of people in a defined location.

Crowd assessment using Wi-Fi signals has been previously used to monitor pedestrians in various environments such as large public events used in mobility simulations [2], university areas for flow characterization [2, 5] and urban environments to build predictive models for trajectories [10]. Furthermore, Wi-Fi *Management Frames* were extensively exploited in occupancy estimation inside office environments [15] and public transport vehicles [9], while also enabling the study of social relationships between mobile users [1, 6]. Likewise, Wi-Fi received signal strength (RSS)[3, 4, 12] and channel state information (CSI)[11, 16] can be largely used in indoor localization and people counting.

In this work, we revisit the use of Wi-Fi signals to estimate crowd size in various environments while preserving the privacy of the monitored subjects. Our main contribution towards this ongoing work is finding very strong correlations between RSS values from Wi-Fi signals sent by nearby fixed devices and the presence of active mobile devices in a large monitored area. These results enable the estimation of

crowd size in a university building without compromising the privacy of the monitored subjects.

2 METHODS

We present our approach for estimating crowd size based on RSS values. Additionally, we discuss the commonly used Wi-Fi *Management Frames*. This type of frame is sent by wireless devices, often containing their unique MAC address and no data payload. From these frames, an altered version of the device identifiers is stored, allowing us to get a zero-th order approximation of the crowd size in order to validate the observations of the aforementioned RSS-based method.

2.1 RSS-Based Estimates

In order to estimate the size of a crowd in a confined space without continuously tracking all device identifiers, and instead using the received signal strength from fixed wireless devices, we explore the signal attenuation caused by the presence of people in a large. We discuss basic signal propagation models which will help us understand the reasoning behind this approach, and will lead us to the fixed sender model discussed (see Section 2.2). Note that estimating the coefficients of the path loss models however, is beyond the scope of this work.

The simplest propagation model for wireless path loss (Ω) is the **free-space path loss** (FSPL) which takes only the distance between sender and receiver and the wavelength of the signal. This model characterizes the path loss as an inverse squared power-law $\Omega \propto d^{-2}$ of the distance d to the receiver. The FSPL for a given distance d and wavelength λ is estimated with Equation 1 [13]:

$$\Omega(d) = 20 \log_{10} \left(\frac{4\pi d}{\lambda} \right) \tag{1}$$

Already from Equation 1 we can expect a Wi-Fi signal at 5 GHz ($\lambda \approx 6$ cm) to have a higher path loss than at 2.4 GHz ($\lambda \approx 12.5$ cm) for the same distance. This basic model can also provide us with an upper-bound of the expected received signal strength, given a fixed infrastructure with nodes at known distances and the transmission power.

Furthermore, the **log-distance path loss** (LDPL) model captures the attenuation as a modified power-law with log-normal variability. The estimation of the total path loss (L_{total}) using this model is given by Equation 2:

$$L_{\text{total}} = \Omega(d_0) + \gamma \log_{10} \left(\frac{d}{d_0} \right) + X_\sigma \tag{2}$$

where d is the distance to the receiver, d_0 is a reference distance to the sender (usually 1 meter), γ is the attenuation exponent, and X_σ is a log-normally distributed Gaussian random variable with zero mean and standard deviation σ. The PL function is often the FSPL or determined experimentally.

In the LDPL model the γ exponent is proportional to the complexity of the path between sender and receiver. Therefore it is defined by the size and materials of physical obstacles along the signal paths, and it is often determined experimentally. In this work, we assume that γ will only *vary* due to the presence of human bodies and their activities, given a pair of fixed sender and receiver in a confined space.

The effect of human bodies on electromagnetic fields is often modeled by its absorption cross section (ACS, σ_a) [8], given by the ratio $\sigma_a = \frac{P_{\text{abs}}}{S_{\text{inc}}}$, of absorbed power P_{abs} and incident power density S_{inc}. This metric is affected by the frequency, direction and polarization of the incident electromagnetic signal, as well as the absorption rate, area and mass of a person's body, and its estimation has applications in wireless infrastructure planning. In our work, we assume it to be approximately the same for every subject, therefore the effect of bodies from different persons is assumed to be constant (*i.e.,* one added person adds a fixed contribution to the TPL with his/hers ACS).

Furthermore, interference from other RF signals at the same frequency can significantly influence the total path loss and make it even more complex to be modelled. For example, when sending signals at 2.4 GHz between two measurement probes[1], the measured RSS may suffer from interference from Bluetooth headphones nearby, which also uses the 2.4 GHz ISM band, as well as any other wireless device using Wi-Fi in the same channel. Therefore, the various activities performed by the subjects present in our studied space can significantly affect the total path loss of a Wi-Fi signal, and we assume this effect in the RSS measures to be proportional to the number of present subjects (*i.e.,* one added person adds a fixed contribution to the TPL with a fixed level of activity).

In this work, we use this understanding of wireless signal attenuation caused by humans and their activities to estimate their group size. Next we present our basic model for the relationship between signal path loss and number crowd size.

2.2 Fixed Sender Model

If we define the path loss (*PL*) of a wireless signal as the difference between transmitted $\mathcal{P}_{\text{tx}}$ and received power $\mathcal{P}_{\text{rx}}$, then PL between a fixed sender *tx* and a fixed measuring probe *rx* in an empty building is $\text{L}_{\text{empty}} = \mathcal{P}_{\text{tx}} - \mathcal{E}_{\text{rx}}$, where $\mathcal{E}_{\text{rx}}$ is the received signal strength from *tx* when the building

[1] A Raspberry Pi with synchronized clock, running a custom software to capture Wi-Fi *Management Frames*.

is *empty*, and it is defined as the maximum RSS for *tx* during all our measurements (*i.e.*, $\mathcal{E}_{\text{rx}} = \max \mathcal{P}_{\text{rx}}$).

The path loss once we have a given number of people in our monitored area is $L_{\text{empty}} + L_{\text{people}} = \mathcal{P}_{\text{tx}} - Q_{\text{rx}}$, where Q_{rx} is the RSS from *tx* at any given time.

This allows us to write $L_{\text{people}} = \mathcal{E}_{\text{rx}} - Q_{\text{rx}}$, therefore eliminating the need to know the transmission power $\mathcal{P}_{\text{tx}}$. Note that this formulation assumes the sender to be fixed (or *stationary*).

In this study, we pose (and validate) the hypothesis that the average L_{people} from all nearby stationary devices is affected by (or, is proportional to) the total number of people inside our monitored building, or simply $\text{PL}/\langle\text{PL}\rangle \propto N_d$. The expected value for *PL* is therefore, given by Equation 3:

$$\text{PL} = \int L_{\text{people}}\ f(L_{\text{people}})\ \text{d}L_{\text{people}} \tag{3}$$

where $f(L_{\text{people}})$ is the density function for a measured value of L_{people}, which is proportional to the log-normal distribution from Equation 2.

For a zero-th order approximation of the number of people in the monitor building, we use Wi-Fi Management Frames, presented in the next section.

2.3 IEEE 802.11 Management Frames

On top of the physical layer, the IEEE 802.11 standard specifies three different frame types on the data link layer: *Data Frames*, which carry the data payload, *Control Frames*, which allow devices to control access to the medium (*e.g.*, Request to Send and Clear to Send frames), and *Management Frames*, which assist Wi-Fi enabled devices in finding and connecting to a wireless network. From *Management Frames*, beacons, null frames, probe requests and probe responses are of particular interest to this work for containing no network traffic information (*e.g.*, browsing data) and being continuously sent by wireless devices.

Wireless devices constantly scan for available access points, gathering information (*e.g.*, network name (SSID), RSS and security configurations) which is then used to choose which network to join. When scanning for nearby networks, devices can do this passively, by waiting for beacon frames sent by access points, or actively, by sending probe requests and waiting for probe responses to finally initiate the connection. Furthermore, null frames are sent by clients about whether it is in an active power state or not.

In this study we use these frames as signals to sample the presence of nearby wireless devices. To counteract the possibility of tracking devices based on frames, different wireless device manufacturers have recently developed MAC randomization techniques[7, 14]. Since MAC address randomization is not yet well understood for all manufacturers, which could lead to a miscount of nearby devices, we discard all randomized addresses from any of our analysis. Furthermore, crowd size estimate based on counting devices from *Management Frames* may lead to inaccurate results due to a variable number of devices per person.

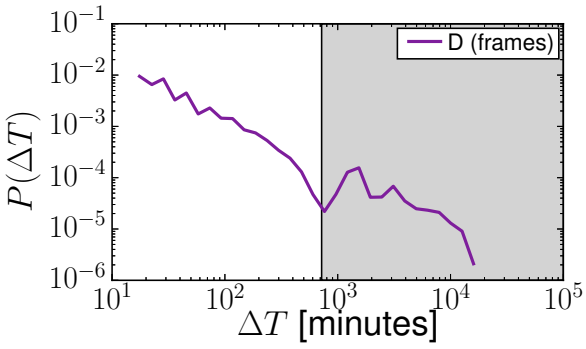


Figure 1: Stay duration with a bimodal distribution, with stationary devices in the gray shaded area.

While estimating crowd sizes in a confined space is possible with an ad-hoc setup [4], achieving similar results in a large area with the same method would require a large number of probes being deployed, and hence requires a different approach. In the next section, we discuss an alternative RSS-based approach which uses signals from fixed (or stationary) devices instead.

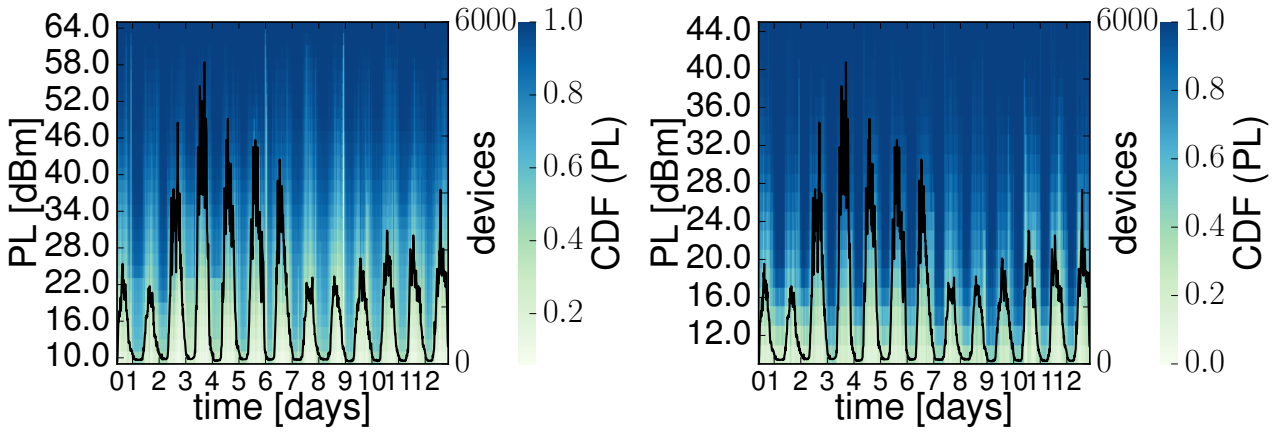
3 FIXED INFRASTRUCTURE EVALUATION

We now apply our RSS-based method to a larger space to get a close estimate on the number of devices. Our method uses the RSS path loss from *Management Frames* sent by stationary devices already available at the building. Next, we explain the measurement setup and discuss our results in which we observe a strong linear relationship between the average path loss and the amount of mobile devices counted from *Management Frames*.

3.1 Management Frames Collection Setup

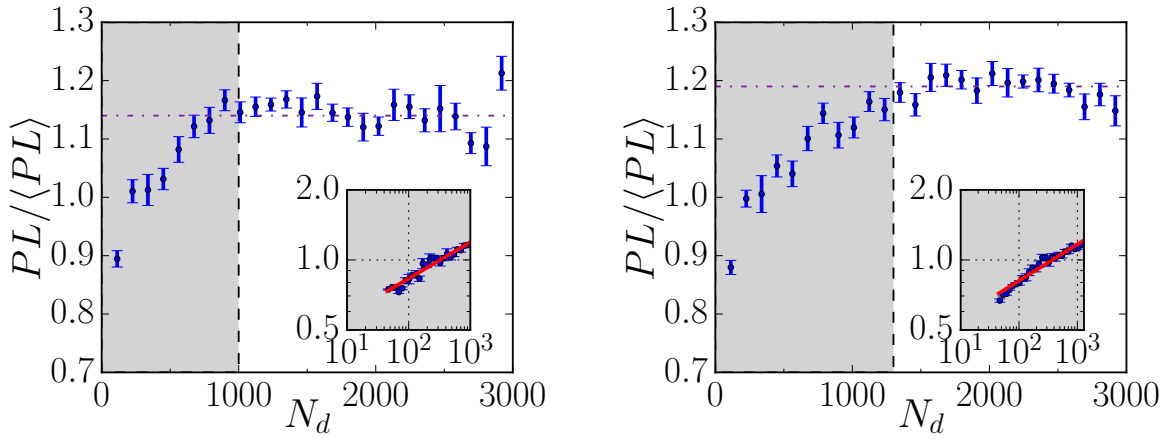
During 13 consecutive days in February/2019, we deployed 4 probes in the Computer Science and Mathematics building of our university, covering the main entrances, a parking lot and the main hall with a cafeteria. This building was chosen for its large size as well as large number of visitors, with over 6000 students and 400 staff members [2] [3]. The building is located in an area exclusively used by the university and

[2] http://www.in.tum.de/en/cover-page/
[3] https://www.ma.tum.de/en/department/about-us.html



(a) 2.4 GHz (b) 5 GHz

Figure 2: Average path loss (PL) and total mobile devices at different frequencies.



(a) 2.4 GHz (b) 5 GHz

Figure 3: Change in average path loss (PL/⟨PL⟩) with the total number of nearby devices N_d. An observed monotonic increase (shaded areas), followed by a saturation in the observed path loss. Insets show the shaded areas well approximated by PL/⟨PL⟩ $\sim N_d^{\alpha}$, with $\alpha = 0.156$ for both frequencies.

other research centers. We will also refer to these location records as the *frames* data set.

For this setup, every probe used 4 external omni-directional antennae capturing frames at channels 6 and 11 (2.4 GHz), 36 and 44 (5 GHz). Once captured, MAC addresses were first classified as *locally* or *globally* managed, or random and non-random respectively. Lastly before storing these records, the device identifiers were hashed with a one-way hash function.In this way, counting devices was possible while anonymizing these identifiers.

We captured over 28 million frames where half were probe requests and half null frames, from over 35,000 global MAC addresses. The median time present in the records was of 1.1 days (25th-percentile: 51 minutes, 75th-percentile: 7 days, 99th-percentile: 13 days) and the median inter event time per device was of 5 seconds (25th-percentile: 1 second, 75th-percentile: 16 seconds, 99th-percentile: 37 minutes). To exclude any distant pass-by devices we discarded any frame for which the RSS was below -90 dBm.

In order to get a closer estimate of the total number of visitors in the monitored building, we classified observed

devices from *Management Frames* into stationary and mobile. We then compare RSS values from the former with device counts from the latter.

3.2 Device Classification

To classify the observed devices from *Management Frames* we compute the probability function of the stay duration $P(\Delta T)$ for the entire set of global device identifiers (*i.e.*, non-random). We define a location event for a wireless device as a Wi-Fi frame received by one of our probes, and define a *stay* at the monitored building by grouping consecutive location events while also grouping pairs of events (with no other known location in between) which were up to 15 minutes apart. We eliminate from our analysis any *stay* shorter than 15 minutes in order to not bias our results with pass-by devices.

From the distribution of $P(\Delta T)$, we observed a bimodal distribution with a clear inflection at 12 hours, as can be seen in Figure 1. As a result, with our aim of using RSS values from static stations and count the number of mobile devices (non-stationary), we classify device identifiers with any *stay* longer than 12 hours as a stationary device.

3.3 Large Area Results

The average measured RSS path loss (PL) values from stationary devices show a clear periodicity with the total number of devices counted from *Management Frames*, as depicted in Figure 2 for signals at 2.4 GHz and 5 GHz. In these plots, the CDF of PL is color-coded, highlighting its variability in accordance with the total number of nearby devices (N_d).

We analyzed the effect of the number of counted devices N_d in the average path loss from all stationary devices. We observed a monotonic increase in $\mathrm{PL}/\langle \mathrm{PL} \rangle$ when N_d varies up to 1000 at 2.4 GHz and up to 1300 at 5 GHz, from which PL no longer changes significantly. Figure 3 illustrates the increased shadowing effect from an increasing number of devices at both frequency bands (gray area), as well as a saturation of the measured path loss from those points. Furthermore, still on Figure 3, the inset plots shows the shaded area intervals being well approximated by $\mathrm{PL}/\langle \mathrm{PL} \rangle \sim N_d^{\alpha}$, where interestingly, $\alpha = 0.156$ for both frequencies.

Additionally, PL and the log of N_d showed a strong Pearson correlation $\rho_{2.4\,\mathrm{GHz}} = 0.74$ at 2.4 GHz and $\rho_{5\,\mathrm{GHz}} = 0.82$ at 5 GHz for the increasing intervals (with p-value $< 10^{-4}$ for both). As expected, there were no statistically significant correlations observed however, between these numbers for the non-increasing intervals at both frequencies (p-value > 0.25).

3.4 Discussion

The strong correlation between N_d and PL indicates the applicability of using RSS values to estimate crowd size. However, the saturation in path loss beyond a certain number of devices at each frequency shows a limitation in the sensitivity of our setup. This limitation suggests that our proposed method depends on the density of the crowd, as well as the level of wireless activity of those individuals (see Section 2.1).

Furthermore, PL varying as the power of the number of devices could be explained by the result of additive log-normally-distributed processes [17], such as the log-distance path loss which models the path loss between sender and receiver (see Section 2.1).

4 CONCLUSIONS

In this paper we presented a privacy-preserving solution to crowd density estimation using Wi-Fi signals. We evaluated the use of received signal strength from Wi-Fi signals in a real-world scenario for these estimates. In a large area, Wi-Fi RSS path loss values from an existing fixed infrastructure were compared to the number of nearby devices.

At both frequency bands, we obtained a strong linear relationship between the presence of over 35,000 mobile devices, using a series of *Management Frames* collected over 13 days, and RSS path losses from signals sent by stationary devices in a large university building. These initial results support the applicability of this method for estimating crowd sizes in large areas.

We plan to extend these results with a denser deployment of probes, coverage of all channels, as well as to incorporate other types of Wi-Fi frames. Furthermore, we will build crowd size estimates from ground truth while also validating the obtained parameters and their dependencies on the type of space we are monitoring.

The present paper illustrates a privacy-preserving alternative for crowd size estimates. Existing methods based on Wi-Fi *Management Frames*, aside from challenges such as MAC randomization, difficult spatial segmentation and multiple devices per person, may compromise the privacy of the monitored subjects. With rising concerns in data privacy, we demonstrate that RSS-based assessments can accurately estimate densities in larger areas without compromising the privacy of the monitored subjects.

REFERENCES

[1] Marco V Barbera, Alessandro Epasto, Alessandro Mei, Vasile C Perta, and Julinda Stefa. 2013. Signals from the crowd: uncovering social relationships through smartphone probes. In *ACM IMC*.

[2] Bram Bonné, Arno Barzan, Peter Quax, and Wim Lamotte. 2013. WiFiPi: Involuntary tracking of visitors at mass events. In *IEEE WoWMoM*.

[3] Krishna Chintalapudi, Anand Padmanabha Iyer, and Venkata N Padmanabhan. 2010. Indoor localization without the pain. In *ACM MobiCom*.

[4] Saandeep Depatla and Yasamin Mostofi. 2018. Crowd counting through walls using wifi. In *IEEE PerCom*.

[5] Yuki Fukuzaki, Masahiro Mochizuki, Kazuya Murao, and Nobuhiko Nishio. 2014. A pedestrian flow analysis system using Wi-Fi packet sensors to a real environment. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM.

[6] Hande Hong, Chengwen Luo, and Mun Choon Chan. 2016. Socialprobe: Understanding social interaction through passive wifi monitoring. In *Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. ACM.

[7] Jeremy Martin, Travis Mayberry, Collin Donahue, Lucas Foppe, Lamont Brown, Chadwick Riggins, Erik C Rye, and Dane Brown. 2017. A study of MAC address randomization in mobile devices and when it fails. *Proceedings on Privacy Enhancing Technologies* (2017).

[8] Gregory C.R. Melia, Martin P. Robinson, Ian D. Flintoft, Andrew C. Marvin, and John F. Dawson. 2013. Broadband measurement of absorption cross section of the human body in a reverberation chamber. *IEEE Transactions on Electromagnetic Compatibility* 55, 6 (2013), 1043–1050.

[9] Lars Mikkelsen, Radoslav Buchakchiev, Tatiana Madsen, and Hans Peter Schwefel. 2016. Public transport occupancy estimation using WLAN probing. In *8th International Workshop on Resilient Networks Design and Modeling*. IEEE.

[10] ABM Musa and Jakob Eriksson. 2012. Tracking unmodified smartphones using wi-fi monitors. In *Proceedings of the 10th ACM conference on embedded network sensor systems*.

[11] Sameera Palipana, Piyush Agrawal, and Dirk Pesch. 2016. Channel state information based human presence detection using non-linear techniques. In *Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments*. ACM.

[12] Moustafa Seifeldin, Ahmed Saeed, Ahmed E Kosba, Amr El-Keyi, and Moustafa Youssef. 2012. Nuzzer: A large-scale device-free passive localization system for wireless environments. *IEEE Transactions on Mobile Computing* (2012).

[13] John S Seybold. 2005. *Introduction to RF propagation*. John Wiley & Sons.

[14] Mathy Vanhoef, Célestin Matte, Mathieu Cunche, Leonardo S Cardoso, and Frank Piessens. 2016. Why MAC address randomization is not enough: An analysis of Wi-Fi network discovery mechanisms. In *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security*.

[15] Edwin Vattapparamban, Bekir Sait Çiftler, Ismail Güvenç, Kemal Akkaya, and Abdullah Kadri. 2016. Indoor occupancy tracking in smart buildings using passive sniffing of probe requests. In *IEEE ICC*.

[16] Wei Xi, Jizhong Zhao, Xiang-Yang Li, Kun Zhao, Shaojie Tang, Xue Liu, and Zhiping Jiang. 2014. Electronic frog eye: Counting crowd using wifi. In *IEEE INFOCOM*.

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

Publication 4

©2021 Springer Nature, reprinted with permission from:

Tonetto, L., Lagerspetz, E., Yi Ding, A. et al. The mobility laws of location-based games. EPJ Data Sci. 10, 10 (2021). https://doi.org/10.1140/epjds/s13688-021-00266-x

RightsLink

The mobility laws of location-based games

Author: Leonardo Tonetto et al
Publication: EPJ Data Science
Publisher: Springer Nature
Date: Feb 15, 2021

Copyright © 2021, The Author(s)

Creative Commons

This is an open access article distributed under the terms of the Creative Commons CC BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.

To request permission for a type of use not listed, please contact Springer Nature

© 2024 Copyright - All Rights Reserved | Copyright Clearance Center, Inc. | Privacy statement | Data Security and Privacy | For California Residents | Terms and ConditionsComments? We would like to hear from you. E-mail us at customercare@copyright.com

Title: The Mobility Laws of Location-Based Games
Authors: **Leonardo Tonetto** (TUM), Eemil Lagerspetz (Univ. of Helsinki), Aaron Yi Ding (TU Delft), Jörg Ott (TUM), Sasu Tarkoma (Univ. of Helsinki), Petteri Nurmi (Univ. of Helsinki)
Journal: EPJ Data Science (10-1)
Publishing date: February 3, 2021
Reference: [5]

Publication Summary

In this article, we explore the impact of location-based gaming, using PokémonGO as an example, on human mobility and the underlying laws governing it. The study analyzes two datasets: one from smartphone application logging and the other from location-tagged social media. The results show that location-based games increase mobility, primarily by encouraging individuals to explore their local neighborhoods more thoroughly. The increase in mobility can be explained by a larger number of short-hops rather than actively visiting new areas. The characteristics of mobility patterns, such as the radius of gyration and the number of visited locations, remain consistent over time regardless of game usage.

The findings suggest that location-based gaming has a measurable effect on physical activity, as it increases daily displacements and encourages individuals to explore familiar regions more thoroughly. However, the study reveals that the increase in mobility does not result from individuals moving to and exploring new regions but rather from changes in everyday activity and exploration within familiar areas. The analysis of the spatial distribution of mobility indicates that the geographic distribution becomes more homogeneously distributed for users with a high anisotropy, supporting the idea that location-based gaming affects mobility patterns.

Overall, our research provides insights into the impact of location-based gaming on human mobility and contributes to our understanding of the factors governing variations in personal mobility. By studying the effects of location-based games on mobility laws, we shed light on the potential benefits of such games in promoting physical activity and improving our collective understanding of human mobility.

Contribution

I came up with the idea with supervision from Jörg Ott and Sasu Tarkoma, designed the study and analyzed the data. The data were originally collected by Eemil Lagerspetz. I wrote the text with support from Petteri Nurmi. All authors reviewed the text.

Tonetto et al. *EPJ Data Science* (2021) 10:10
https://doi.org/10.1140/epjds/s13688-021-00266-x

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

The mobility laws of location-based games

Leonardo Tonetto[1*], Eemil Lagerspetz[2], Aaron Yi Ding[3], Jörg Ott[1], Sasu Tarkoma[2] and Petteri Nurmi[2]

*Correspondence:
tonetto@in.tum.de
[1]Techinical University of Munich, Boltzmannstrasse 3, 85748 Garching bei München, Germany
Full list of author information is available at the end of the article

Abstract

Mobility is a fundamental characteristic of human society that shapes various aspects of our everyday interactions. This pervasiveness of mobility makes it paramount to understand factors that govern human movement and how it varies across individuals. Currently, factors governing variations in personal mobility are understudied with existing research focusing on explaining the aggregate behaviour of individuals. Indeed, empirical studies have shown that the aggregate behaviour of individuals follows a truncated Lévy-flight model, but little understanding exists of the laws that govern intra-individual variations in mobility resulting from transportation choices, social interactions, and exogenous factors such as location-based mobile applications. Understanding these variations is essential for improving our collective understanding of human mobility, and the factors governing it. In this article, we study the mobility laws of location-based gaming—an emerging and increasingly popular exogenous factor influencing personal mobility. We analyse the mobility changes considering the popular PokémonGO application as a representative example of location-based games and study two datasets with different reporting granularity, one captured through location-based social media, and the other through smartphone application logging. Our analysis shows that location-based games, such as PokémonGO, increase mobility—in line with previous findings—but the characteristics governing mobility remain consistent with a truncated Lévy-flight model and that the increase can be explained by a larger number of short-hops, i.e., individuals explore their local neighborhoods more thoroughly instead of actively visiting new areas. Our results thus suggest that intra-individual variations resulting from location-based gaming can be captured by re-parameterization of existing mobility models.

Keywords: Human mobility; Mobile applications; Location-based games

1 Introduction

Location-based gaming has steadily emerged as a popular pastime on smartphones, and become a potentially effective way at promoting physical activity [1–3]. From a scientific standpoint, the most unique and interesting aspect of these games is how they encourage and promote movement, which can improve physical and mental health [4, 5], and be comparable to a health or a fitness app [1–3, 6, 7]. More generally, location-based games are examples of a broader class of smartphone applications that attempt to promote physical activity—either directly through recommendations or indirectly through objectives that are linked with physical locations [1–3, 8]. Other examples of applications in this cat-

© The Author(s) 2021, corrected publication 2021. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Springer

egory include varied location-based services [9], online location-based social networks [10] and smartphone and wearable applications for physical activity [11]. In this article, we focus on location-based games due to their immense popularity and their use of gamification, which has been shown to be among the most effective mechanisms for achieving sustained change in mobility [12, 13]. We study mobility changes in response to location-based gaming through Pokémon GO, the best known, and one of the most popular examples of location-based games. Pokémon GO remains among the most popular mobile apps in many countries, it has over 100 million active users, and has been downloaded over billion times in total. Pokémon GO is not an isolated success story either with other location-based games, such as `Zombies`, `Run!`, `Ingress`, `Geocaching`, `Minecraft Earth` and `Harry Potter: Wizards Unite` similarly being highly popular.

Current research understanding suggests that location-based games, and related applications, can have a sustained effect on mobility. Indeed, several studies have demonstrated smartphone applications to have an effect on the daily activity levels of their users [1, 8, 14]. As an example, empirical studies based on pedometers have demonstrated that Pokémon GO has an effect on mobility, resulting in an increase of around 1400 steps for each day that the user plays the game, and the total effect lasting for at least 30 days [1]. Similar findings have also been obtained from quantitative analyses, e.g., Colley et al. [15] characterized players through questionnaires and geostatistical analysis of game elements and highlighted that Pokémon GO may have introduced significant changes to their mobility. These studies, however, have also shown that the retention of the application and the effect on physical mobility tends to be short-lived with persuasion mechanisms, such as gamification, and social interactivity, being central to prolonging the positive effect on mobility.

While the overall effect on physical activity and movements has been established, important gaps still exist in our understanding. In particular, little information currently exists on *how* this increase affects the characteristics of the user's mobility patterns or *which factors* mediate these effects, and more importantly, how these changes affect the underlying laws governing human mobility. Indeed, as many of the game mechanisms in Pokémon GO are centered around physical movement, the changes could result from increases in everyday physical activity instead of changes in personal mobility patterns. In this article, we explore how the changes in mobility induced by Pokémon GO relate to the laws governing human mobility, in an attempt to fill up this gap in current scientific understanding. We analyse displacement data captured from two sources to obtain a detailed view of mobility patterns and how they are influenced by Pokémon GO. The first data set (Dataset-A) consists of mobile phone app logging (Carat, an energy-awareness app) from over 3900 users, and the second (Dataset-B) of location-tagged social media (Twitter) from over 21,500 users. The granularity of location information differs in these data sets, with social media providing GPS coordinates and app logging providing coarse grained estimates of total displacements with approximately 2 km resolution. Our longitudinal dataset captures time before, during and after Pokémon GO's initial peak in popularity from January 2016 to June 2017 (18 months), allowing us to better study the duration of the game's effects, as well as to account for potential novelty effects (see the Section Datasets for details about the data, data collection process, and data validity).

The results of analysis show that Pokémon GO does indeed increase mobility—in line with previous findings—but the characteristics governing mobility remain consistent with

a truncated Lévy-flight model and that the increase in mobility can be explained by a larger number of short-hops, i.e., individuals explore their local neighborhoods more thoroughly instead of actively visiting new areas. Our results thus suggest that intra-individual variations resulting from location-based gaming can be captured by re-parameterization of existing mobility models. Besides offering novel insights into variations in personal mobility and contributing to our collective scientific understanding of human mobility, our results have practical implications to transport policy planners (e.g., improve design of on-demand transport networks), epidemiology (e.g., explaining characteristics of mobility patterns and offering insights into potential disease transmission routes), urban sciences. and other fields.

2 Datasets

We analyze mobility through two datasets, one collected by instrumenting mobile phones with the Carat energy-awareness applications, and the other obtained from Twitter. We include data from January 2016 to March 2018 from Carat and January 2016 to June 2017 from Twitter. The target game Pokémon GO was released in July 2016. To assess the impact of Pokémon GO on gamers' daily mobility and to validate the generality of our findings, we make compare Pokémon GO use in each of the two datasets against contrasting but complementary baselines. For Carat we compare Pokémon GO users with players of Clash Royale, a non-location-based game that was one of the most popular games released in 2016, whereas for Twitter we compare *gamers* and *non-gamers* (i.e., *infected* vs. *control-group*).

2.1 Dataset A: Carat

The first major dataset used in this study was collected by application logging integrated as part of the Carat[1] [16] smartphone application. This Android and iOS app collects information from the mobile device it is running on and recommends personalized actions aimed at increasing battery life [17].

Carat uses energy-efficient and non-invasive instrumentation to record the state of the device, including a list of running apps, mobile network technology being used (e.g., WiFi or LTE), and distance traveled since the last record. Each of these values is recorded at every 1% battery level change (either charging or discharging) and it also contains a uniquely identifiable id per user and timestamp. The Carat application does not run on the background, but instead registers to the smart device OS's battery change events. Because of this, Carat's data can miss events that happen when the device is in deep sleep, when the application is evicted from memory by the OS, or when the Carat application has been terminated manually by the user. This results in a temporally sparse dataset that requires preprocessing with suitable statistical methods.

Since its first release in 2012, Carat has been deployed in over one million mobile devices in dozens of countries. For our study, we analyze a subset of these data spanning from January/2016 until March/2018. We consider only Android users as the IOS version of the time no longer supported logging the list of running applications.[2] This subset includes

[1] http://carat.cs.helsinki.fi/

[2] iOS 9.3.4 was released on August 4, 2016: https://www.macrumors.com/2016/08/04/apple-releases-ios-9-3-4-with-security-fix/.

Table 1 Left: Number of gamers on Twitter. Right: Number of gamers on Pokémon GO (PG) and Clash Royale (CR)

Twitter

City (code), Country	N.
São Paulo (SPO), Brazil	924
Jakarta (JKT), Indonesia	911
London (LON), UK	853
Singapore (SIN), Singapore	709
Santiago (SCL), Chile	661
Tokyo (TKY), Japan	631
Bangkok (BKK), Thailand	599
San Francisco (SFO), USA	597
New York (NYC), USA	564
Toronto (TOR), Canada	447
Paris (PAR), France	373
Seattle (SEA), USA	348
Boston (BOS), USA	279
Sydney (SYD), Australia	268
Hong Kong (HKG), China	263
Barcelona (BCN), Spain	247
Moscow (MOW), Russia	143
Helsinki (HEL), Finland	92

Carat

Country (code)	PG	CR
USA (us)	780	134
Finland (fi)	746	175
Germany (de)	495	79
UK (gb)	153	20
Canada (ca)	149	20
India (in)	122	137
Japan (jp)	113	6
Spain (es)	102	58
Italy (it)	78	43
Netherlands (nl)	50	9

173.6 million records from 74,000 users out of which 3996 played the game at least once on Android. We classify a Carat user as a *gamer* from his/her first record containing Pokémon GO as a running application.

To identify the effect of Pokémon GO on mobility, in our analysis of the Carat dataset we compare the effects of Pokémon GO and Clash Royale—a non-location-based game—on their players. We ensure these gamers had records before the day of the installation of the respective app as well as records after the last day it was observed in our records. Released in March/2016, Clash Royale[3] is a multi-player game in which users battle in support of their clans. Fundamentally, while Pokémon GO requires its users to physically move to reach other players and in-game objects, Clash Royale is agnostic to any sensor in the phone and allows any two or more players to interact regardless of their location. We study a total of 1323 users who played this game at least once on Android. Table 1 lists the number of users of these two games in the Carat dataset (for the top 10 countries).

For every new Carat record, the app stores the geographical coordinates of the device locally. For that, it queries the coarse-grained last known location from the Location Manager API for Android. The individual measurements are not stored by Carat, and only distance between consecutive records is transferred to the back-end for further analysis and location information from older records is destroyed. The benefits of this approach are twofold: lower battery consumption since it does not use the power-hungry GPS chip of the device, and the privacy of the user is preserved by never disclosing the exact location of the user. One limitation of this method is the variable accuracy of these location services. The distribution of displacements from Carat shows an abrupt inflection (knee) at around 2 km. This may be due to Android's coarse-grained location granularity, which mostly seems to report location with a cell-tower precision (around 2 km accuracy).

[3] https://clashroyale.com

2.2 Dataset B: Twitter

As our second dataset, we analyse 8.7 million geotagged tweets from over 21,500 users in 15 different countries. Studying this diversity of countries allowed us to mitigate the impact of possible regional bias in our analyses. To obtain these records, we first queried Twitter's web page following a certain criterion, resulting in a list of users. From each user account in this list, we downloaded its entire *timeline* (set of tweets) through Twitter's REST API,[4] keeping only those records with a geotag (17.4% of the total). For both *gamers* and *non-gamers* the query criteria were (i) a given location (e.g., Bangkok, Thailand), and (ii) a period within the time of our study. This approach ensures the availability of these data for reproducibility, as the access to Twitter's REST API[5] is the only requirement. Furthermore, Twitter's developer policy precludes long-term storage of location-based data.

For *gamers*, we require the string *#PokemonGo* to appear in the tweet (or some capitalization alternatives, e.g., *#pokemongo*). We collected tweets from over 8900 *gamers*. Manual inspection of 1% randomly sampled tweets from this set revealed the content of the tweet to be associated with the game in 90% of the cases (e.g., screenshots or text about in-game actions), in line with the measurements of Althoff et al. which were based on queries from a web search engine [1]. To eliminate unwanted noise from bots in our Twitter set, we used the *Botometer* [18] API and identified 3.1% of such profiles, which were then discarded. The list of cities included in this study, along with the corresponding total number of *gamers* is shown in Table 1.

The average number of tweets for *gamers* and *non-gamers* was statistically similar (mean (μ): 390 vs. 351, median: 200 vs. 159, probability of the two distributions being identical $p < 0.001$) as well as the inter-arrival-time of tweets per user (in hours, μ: 57.8 vs. 58.1, median: 7.68 vs. 8.13, $p < 0.001$). In all cities analyzed, these geo-tagged tweets were similarly distributed in space among both groups, with urban areas resulting in higher densities of tweets.

2.3 Supporting dataset—Google trends

To define the periods of highest activity of a game, we compared some of the aforementioned metrics with the Google Trends index[6] of *Pokémon GO* as a search term (G). This metric measures the popularity of a search term, with values ranging from 0 (lowest) to 100 (highest). It allows us to validate the *trendiness* of the game per country over a period of time.

2.4 Dataset validity

The combination of datasets used in our work gives us insights on various aspects of how mobile location-based games influence human mobility. Our two main datasets cover large amounts of users during periods before the release of the studied application, and the months during which it had its highest popularity in various countries. Carat's longitudinal dataset contains fine grained measurements about users' displacements and app usage, enabling the study of the impact of the game on mobility as well as investigating

[4] https://developer.twitter.com/

[5] https://developer.twitter.com/en/developer-terms/agreement-and-policy

[6] https://trends.google.com/

various aspects of game retention, such as the availability of cellular networks and battery consumption. Twitter, in turn, allows us to quantify the impact of the game on users' visited areas by labeling those discussing the game as *gamers*.

Carat records are captured without user intervention but contain only displacements between samples. Twitter records contain a precise geographical location but their availability is subject to the user's desire to share the information. These different characteristics allow us to study complementary aspects of the effect Pokémon GO has on its gamers which would not be possible through a single source.

3 Methods

In this section, we describe the methods and metrics used to study the datasets described in Sect. 2. Table 2 lists the probability functions $P(x)$ of the distributions used to model our data.

3.1 Spatial clustering

Given the strong urban aspect of the game and small range of distances traveled while playing it (<10 km, <6.2 mi) [15, 19], we applied a series of clustering algorithms to identify which records are from the user's normal geographic area. Specifically, we classify Twitter records as *local* or *away* depending on their distance from the user's city. These labels were computed with respect to the city from which a user was initially discovered. Since users may visit other cities and countries which may be many kilometers *away*, we focus our study in the *local* area of each city and discard all samples labeled as away.

We classify a given city C using a two-stage clustering process. Let (C_S) denote all of the geographical coordinates of C regardless of user-id. We first cluster C_S using DBSCAN [20] with $\epsilon = 2$ km (maximum distance for two points to be in the same cluster). We then calculate the center of mass C_{cm} of the cluster with the most records and compute the distances between every point in C_S and C_{cm}, namely $d_{s,\text{cm}}$. Finally, we cluster the log transformation of these distances ($\rho = \log(d_{s,\text{cm}})$) using KMeans with $k = 2$clusters (number of clusters the algorithm should look for). These algorithms were chosen for their simplicity and their characteristics making them well-suited for the two phases. DBSCAN is a density-based clustering algorithm that allows grouping points based on a maximum distance, whereas KMeans offers control for the number of clusters to extract in the second phase.

The resulting probability distribution of ρ showed a consistent separation between the *local* and *away* clusters at around $\rho = 5$ (100 km or 62 mi) in all 18 cities studied. We conjecture that this is the typical maximum commuting distance a person would regularly travel, regardless of geographical location. From the classified records, for a given city C, we studied the trajectories of *local* tweets of users who have at least 25% of their records at C.

Table 2 Probability functions for different distributions

Distribution	Probability function $P(x)$
Power law	$x^{-\alpha}$
Truncated power law	$x^{-\alpha}e^{-\lambda x}$
Exponential	$e^{-\lambda x}$
Stretched exponential	$x^{\beta-1}e^{-\lambda x^{\beta}}$
Log-normal	$\frac{1}{x}\exp[-\frac{(\ln x-\mu)^2}{2\sigma^2}]$

3.2 Place extraction

For analysing mobility using displacement data, we need to identify hops that correspond to successive trajectories of users. We accomplish this using an approach that is motivated by algorithms designed for extracting significant locations from sequence data [21]. First, we define a *stop* as a sequence of records following three rules: (1) no displacements are observed, (2) intervals between samples are shorter than a threshold ($\Delta T < \tau$), and finally (3) the sequence spans a minimum amount of time (also τ for simplicity). Furthermore, we define a *movement* following rule (2) as well as being interrupted by any *stop*. To further benefit from Carat's faster sampling rate, for our analysis we require a *movement* to be immediately preceded (also within a max interval τ) by a *stop*. This approach significantly decreases the uncertainty about when a movement actually started and allows us a more accurate view of the users' mobility. For this analysis, we set τ to 15 minutes, allowing us to capture *stops* of that duration while accounting for sampling bias as well as very shorts stop along the way (e.g. traffic lights) and ensuring *movements* are more likely to start from important locations the user might visit.

3.3 Temporal analysis

Figure 1 shows the level of *Pokémon GO* activity in the second half of 2016, through 4 different metrics: number of installations (I, Carat), number of gaming sessions (S, Carat), number of tweets with *#pokemongo* (N_t), and the Google search index for *Pokémon GO*. The shaded area in all plots highlights the highest levels of activity, between July/2016 (release of the game) and end of September/2016. For the analysis of the Twitter dataset, we therefore define 3 equally long periods for our study: *before* the game (April–June/2016), *during* (July–September/2016) and *after* (October–December/2016).

Since we are analyzing the overall impact of the game on *gamers*' mobility, and comparing it with *non-gamers*, we only consider users who had records in all three of the periods as well as *gamers* whose first game activity (e.g., tweet containing *#pokemongo*) was between beginning of July and end of September/2016.

To ensure the validity of our results, we compare observations across the two datasets. From Twitter, the average time between the first and last tweet containing *#pokemongo* is 59.2 days (median: 34.6 days, σ: 69.2 days), significantly smaller than the number of days gamers were observed playing Pokémon GO on Carat (99 days). Despite this difference, both present very similar power law exponents for the distribution of these reletable time intervals (Carat: $\alpha = 1.285$, Twitter: $\alpha = 1.305$).

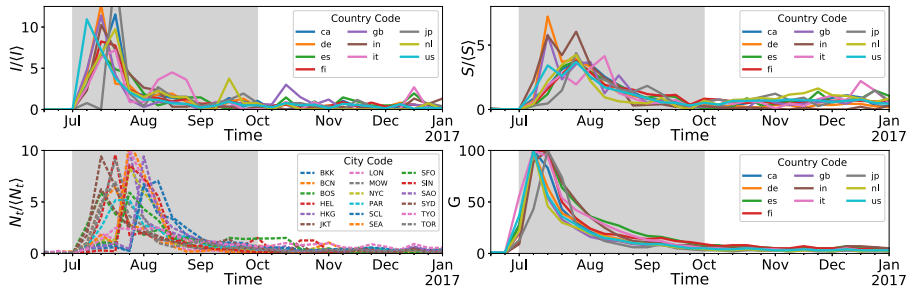


Figure 1 Time series of Pokémon Go activity from new installations (I), game sessions (S), number of tweets (N_t), and Google Trend index (G). The first three are normalized by their average ($\langle\bullet\rangle$)

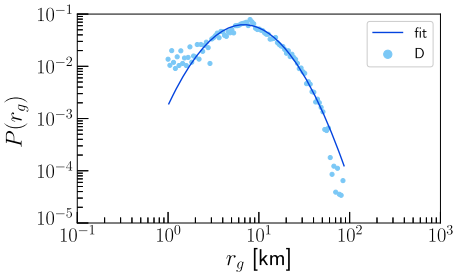


Figure 2 Distribution of radius of gyration $P(r_g)$ for Twitter *D*, best approximated by a log-normal distribution (fit)

3.4 Radius of gyration (r_g)

This commonly used mobility metric [22] conveys the size of the dispersion of a user's studied trajectories. It can be interpreted as the radius of a circle covering the trajectories of a user, centered at the center of mass of all observed points. A gamer that moves to new areas would thus have an increased r_g, but it would remain the same for a gamer playing in the same area. The radius of gyration r_g is calculated with Equation (1):

$$r_g(t) = \sqrt{\frac{1}{n_c(t)} \sum_{i=1}^{n_c} (\overrightarrow{r_i} - \overrightarrow{r_{\text{cm}}})^2}, \tag{1}$$

where $\overrightarrow{r_{\text{cm}}} = \frac{1}{n_c} \sum_i^{n_c} \overrightarrow{r_i}$ represents the center of mass of all visited locations by a given user, and $\overrightarrow{r_i}$ represents location $i = 1, \ldots, n_c(t)$ up to time t.

For the Twitter data set, Fig. 2 depicts the probability distribution $P(r_g)$ and the corresponding best fit model of a lognormal distribution (i.e., $\ln(r_g)$ is normally distributed). Given that our analysis is constrained to *local* points, we therefore also limit trip lengths (i.e., we study regular flights of less than 100 km in the studied cities). Under similar constrained circumstances, a lognormal distribution has been observed by Zhao et al. [23]. Both user groups had very similar distribution parameters, $\mu = 2.44$ and $\sigma = 0.705$ for *gamers* and $\mu = 2.43$ and $\sigma = 0.7244$ for *non-gamers*.

The distribution of $P(r_g)$ being well described by a lognormal function implies that this mobility metric is a result of a multiplicative random process [24]. Therefore, we conjecture that the area covered by a user's *local* trajectories is a result of mechanisms such as transport prices, locations of origin and destination and availability of certain means of transportation.

3.5 Isotropy ratio (σ_y/σ_x)

While the radius of gyration r_g describes the size of the area covered by a user's trajectory, isotropy [22] describes how a user's trajectories are dispersed inside this area given a common reference frame (e_x, e_y). For example, a highly anisotropic set of trajectories would have most of its points dispersed along one of these axes and fewer close to its orthogonal axis. This metric allows us to capture changes in the visits of gamers who play in the vicinity of previously visited locations (and whose r_g may not change).

As proposed by Gonzalez et al. [22], using moment of inertia, we calculate the intrinsic reference frame of a user's trajectories (e_1, e_2), then rotate it around its center of mass into a common reference frame (e_x, e_y). Finally, the dispersion of the observed points of a user

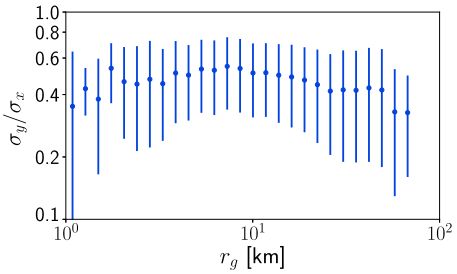


Figure 3 Distribution of Isotropy ratio σ_y/σ_x for different values of radius of gyration (r_g). Average values (dots) and standard deviation (bars)

can be calculated along each axis of this common reference frame. We use the ratio (σ_y/σ_x) of these two since it captures the proportionality of these dispersions along both axes.

For Twitter, Fig. 3 shows the distribution of σ_y/σ_x for varying values of r_g. Given that our analysis is constrained to *local* points, the average ratio observed is higher than in previous works [22, 25], especially for higher values of r_g. This outcome possibly captures the tendency for more isotropic trajectories in urban environments. The behavior was similar for both *gamers* and *non-gamers*.

3.6 Number of visited locations (φ)

To perform a user-centric as well as a location-centric study of visits, we perform spatial binning of the observed tweets. We bin every observed point to the nearest intersection of a mesh grid of 250 meter by 250 meter square cells. Every studied city is covered with this grid. The binning allows us to correct for GPS inaccuracies, as well as group visits which may fall within the area of a large city block.

For Twitter, the distribution of the number of visited locations (φ) between January/2016 and June/2017 is well-described by a stretched exponential (Table 2). For this analysis, we only considered users who were registered before 2016. With a stretching exponent β close to 1, its behavior can be approximated to that of an exponential distribution. This allows us to estimate the number of visits a user will make after some time t by $\varphi(t) = 1/\lambda(t)$, where $\lambda(t)$ is the average number of visits per user, and the value of $\varphi(t)$ will be independent of the users already sampled. If this observation persists for *gamers* while playing, the game would have a different impact on each player's visitation distribution. The distribution fit parameters for *gamers* were $\lambda = 0.0226$ and $\beta = 0.946$, whereas for *non-gamers* we observed $\lambda = 0.0193$ and $\beta = 0.916$. Note the higher average visitation ($1/\lambda$) for *non-gamers*.

3.7 Gaming session (S)

To better understand the behavior of users while playing the game, we define a gaming session (S). Figure 4 depicts the distribution of inter-record-time (ΔT) for all Carat users. A *gaming session* is then defined as a sequence of records (containing the game) in which $\Delta T < 5$ minutes (shaded area in $P(\Delta T)$). A limitation of the Carat dataset is that the application can only record device behavior when running. The background process of Carat may be terminated by the OS at any time when the user is not actively using the Carat app. Therefore the data from Carat are inherently sparse, and records may be missing throughout the day. Given these constraints, for any analysis of S, we only consider those longer than 5 minutes.

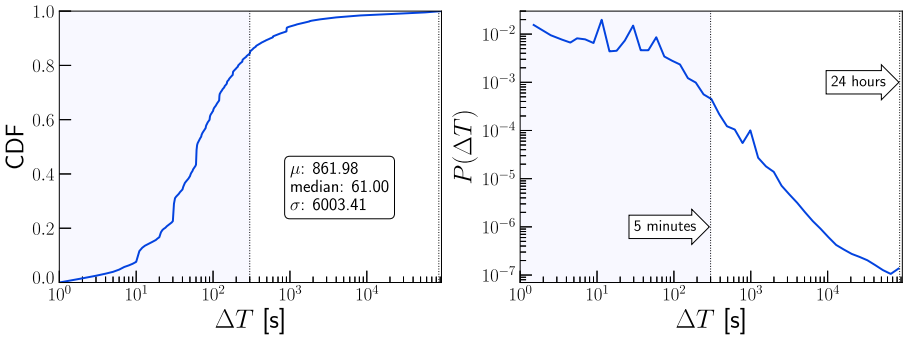


Figure 4 Carat inter-record-time ΔT. Maximum ΔT for two records to be grouped in the same session is 5 minutes (shaded area)

Figure 5 Duration of game sessions for different countries *Pokémon GO*

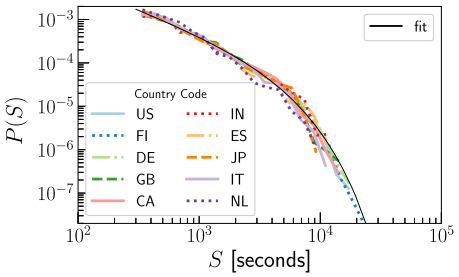
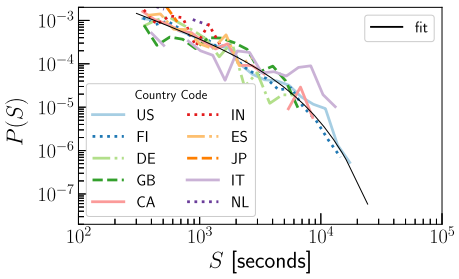


Figure 6 Duration of game sessions for different countries *Clash Royale*



Mobile application usage has been shown to reflect geographic and cultural boundaries [26], which suggests that cultural factors could mediate the results. To demonstrate that this is not the case, and that Pokémon GO usage is highly similar across countries, Fig. 5 and Fig. 6 compare the gaming session times of Pokémon GO and Clash Royale across different countries included in our analysis. From these plots we can observe that the session times for both Pokémon GO and Clash Royale are highly similar across all countries, suggesting that cultural factors have little or no effect on how the games are played. Indeed, the distributions of gaming sessions shown in Fig. 5 and Fig. 6 are consistently similar across the different countries.

3.8 Distance traveled between consecutive records (Δr)

Given the set of (*local*) tweets from a user, Δr is the computed distance between two consecutive records. For simplicity and scalability, this distance is calculated as a straight line between these two points, and not the length of the shortest path between them [27]. Since

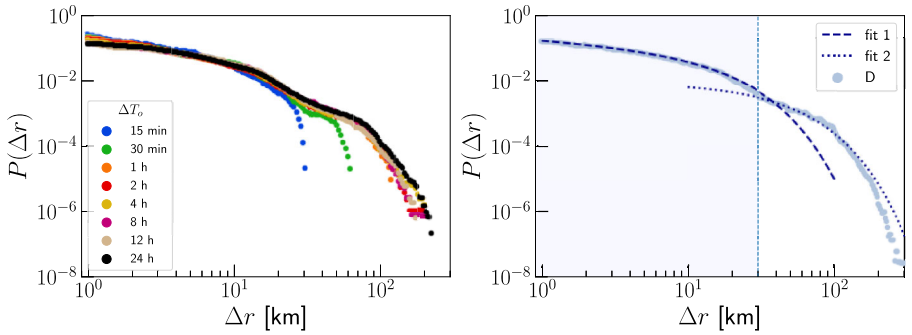
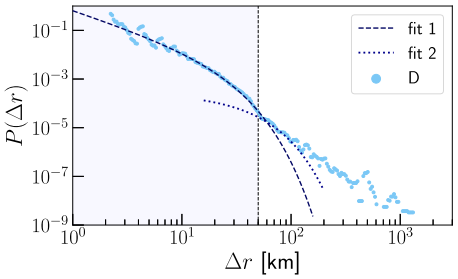


Figure 7 Distribution of displacements $P(\Delta r)$ for a fixed time interval ΔT_o and a multi-modal fit for Twitter *D* by a truncated power law (fit 1) and an exponential (fit 2), split at 30 km. From entire observation period of Twitter

Table 3 Distribution parameters for fits in Fig. 7

Twitter users	Truncated power law (fit 1)	Exponential (fit 2)
Gamers	$\alpha = 0.279, \lambda = 0.089$	$\lambda = 0.036$
Non-gamers	$\alpha = 0.329, \lambda = 0.083$	$\lambda = 0.036$

Figure 8 Distribution of displacements $P(\Delta r)$ for Carat *D* and a multi-modal fit by a truncated power law (fit 1) and an exponential (fit 2), split at 50 km



we are only considering points within an urban environment, we discard all Δr where the corresponding velocity was >120 km/h (75 mph, maximum speed limit on highways).

For Twitter, Fig. 7 depicts the probability distribution $P(\Delta r)$ for a fixed time interval (ΔT_o) as well as for the entire dataset (D). The former shows that $P(\Delta r)$ is not affected by different sampling rates when only *local* records are considered. The latter shows a multi-modal distribution of $P(\Delta r)$, composed by a truncated power law (fit 1) and an exponential (fit 2), divided at an inflection point around 30 km (18.6mi). Similar to results by Jurdak et al. [25], this result validates the multi-modal aspect of human mobility, where short and long distances are covered using different means of transportation. Parameters of each distribution fit had similar values between user groups, summarized in Table 3.

The first part of the model (fit 1) being a truncated power law implies a relative proportionality between a distance traveled and its probability, up to a cut-off point from which probabilities decrease much faster ($\sim 1/\lambda$). Likewise, the second part of the model (fit 2) being an exponential implies that the probability of distances traveled diminishes very fast, rendering very high values of Δr extremely rare.

Since we are not able to distinguish between *local* and *away* points for Carat as we did for Twitter, we limit our analysis of the former to displacements which are smaller than 100 km (62 mi). Similar to our Twitter analysis, Fig. 8 depicts a multi-modal fit for

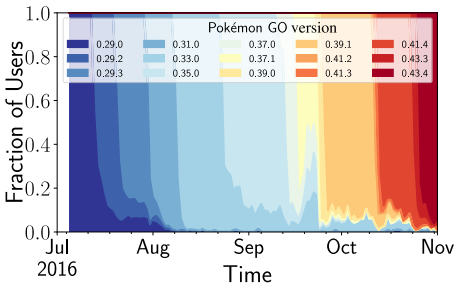


Figure 9 Pokémon GO version adoption (Carat users)

$P(\Delta r)$ with a truncated power law (fit 1) and an exponential (fit 2), although for Carat data, the distribution is split at 50 km (31 mi). Note that Twitter users are more likely to share their location while at their destination, whereas Carat is capable of sampling intermediate displacements. These scaling differences can explain the different exponents of the power laws ($\alpha_{\text{Twitter}} = 0.329$, $\alpha_{\text{Carat}} = 1.469$). It is interesting to note the similarities between datasets in the decay parameters of the exponential cut-off in fit 1 ($\lambda_{\text{Twitter}} = 0.08386$, $\lambda_{\text{Carat}} = 0.08346$) and the exponential in fit 2 ($\lambda_{\text{Twitter}} = 0.03589$, $\lambda_{\text{Carat}} = 0.04505$).

3.9 Performance and usability mediate mobility change

To obtain further insights into factors mediating mobility, we next perform an analysis of how the increase in mobility correlates with different release versions of Pokémon GO. To understand changes in technical functionality, we also correlate our findings against Pokémon GO changelogs.

From Carat's user base, the percentage of adoption for each version during the first four months of its release is depicted in Fig. 9. Changelogs of these early versions point to battery issues being addressed in versions 0.31 and 0.33. Analysis of the expected time a user played the game given the initial version they first played shows a statistically significant increase of 117% (3.5 days to 7.6 days) between these two versions. This result suggests that performance and usability effects mediate mobility changes (and retention). Conversely, our results suggest that location-based games may struggle at achieving persistent change in mobility if they have performance or usability issues.

4 Results

4.1 Location based online game introduces significant changes to mobility

We first validate that Pokémon GO indeed has a significant effect on mobility. To demonstrate this, we split the records in Dataset-A (i.e., Carat) between week-days and week-ends, and categorize users into three groups: low, intermediate, and high engagement users, according to the number of days they were observed playing (A: $[1, 21)$ days, B: $[21, 90)$ days, C: 90 or more days). Separating week-days and week-ends is essential for eliminating possible biases resulting from daily and weekly routines in mobility characteristics [28, 29], whereas categorizing the users is necessary to control for differing engagement levels [30, 31] (see Sect. 3). To control for the effect of location-based game design features, we compare Pokémon GO against Clash Royale, a mobile game without location-based features that was highly popular during the observation period. The average daily displacements calculated from Dataset-A are summarized in Table 4. Statistically significant increases were found for groups B and C (over 2 km and 1 km, respectively)

Table 4 Daily movements (in km), per group according to the number of days playing—A: [1, 21) days, B: [21, 90) days, C: 90 or more days, highlighting statistically significant changes, for Pokémon GO (PG) and Clash Royale (CR). The sample sizes were (995, 1051, 1160) and (257, 317, 230) for (A, B, C) on PG and CR respectively

Game	Period	A	B	C
CR	Week-day	30.3	30.2	32.5
	Week-end	28.7	26.0	28.3
PG	Week-day	27.3 → 31.2	**28.0 ⇒ 29.9**	**30.6 ⇒ 31.6**
	Week-end	29.3 → 29.4	**28.1 ⇒ 30.4**	**29.6 ⇒ 31.4**

when comparing Pokémon GO use to time before it. The increase is significant for both weekdays and weekends ($p < 0.02$). For low engagement users and users of Clash Royale, no statistically significant differences were observed ($p > 0.09$). For groups B and C, the increase in mobility persists even after Pokémon GO use ends.

To validate that the increase in daily displacements is not biased by the use of app logging as a sampling mechanism or the user population of said app, we separately compute the total daily Δr from Dataset-B for all users with at least 3 records per day. We split the users into a `gamer` and a `control` group depending on whether they had used Pokémon GO or not. Similarly to the results for Dataset-A, we observe a statistically significant increase in total daily Δr for gamers during week-days, from 13.1 km to 14.6 km ($p = 0.03$). Conversely, there is a decrease in Δr for `control` group from 16.2 km to 15.9 km ($p = 0.03$). The small, but nevertheless statistically significant, decrease in mobility for the control group is likely explained by a combination of different factors with seasonality and decreased retention, and hence reduced Twitter activity, over time being among the contributing factors. During week-ends, there were no statistically significant differences for `gamers`, but there was a decrease in total daily Δr for `control` group users between the last two periods (16.7 km to 15.2 km, $p = 0.007$).

4.2 Increased mobility from exploring nearby vicinity

The increase in mobility could be explained by three hypotheses: (i) individuals move to and explore new regions, (ii) they explore familiar regions more carefully, or (iii) they engage in higher level of physical activity without exploring any new areas. For example, an increase in step count, could result from increased everyday routine activity instead of changes in personal mobility patterns. As Pokémon GO incorporates several game mechanics that require physical activity from the users to progress and to accumulate achievements with the game, it is essential to separately analyze the extent to which increased mobility affects underlying mobility laws (hypotheses (i) and (ii)) and to which it results from the game mechanics (hypothesis (iii)). To explore the first hypothesis, we use Dataset-B to calculate the evolution of r_g, i.e., the radius of gyration across the different periods. We cluster users by their r_g before, during, and *after* the game at intervals of 5 km, up to 50 km and an additional cluster for $r_g > 50$ km. We observe a strong monotonic relationship in the distributions of r_g between each studied period (for all comparisons: Spearman's rank correlation coefficient $r_s > 0.75$, $p = 0$). However, there were no significant changes in r_g across those months. For both the gamer and the control groups, only those with initial r_g values of 5 km and 10 km showed changes greater than 10%: `gamers`: 7.38 km and 11.18 km, `control`: 7.08 km and 11.36 km, respectively. For all clusters and

observed periods, a statistical test for distribution similarity between `gamers` and `control` had ps > 0.05. The results thus strongly indicate that the geographic area within which users move remains consistent over time regardless of the user playing Pokémon GO or not, i.e., we find no support for the first hypothesis.

To explore the second hypothesis, we assess the total number of locations visited (φ) by users. For `gamers` we observe a small but statistically significant increase during gameplay. Before exposure to the game (April–June/2016), `gamers` visited on average 15.4 locations ($p < 0.001$) whereas for `control` users the respective average number is 18. However, during the game, `gamers` visited two more locations than before (17.4, $p < 0.001$) while for `control`, there was no statistically significant difference in the number of locations visited before and during (18.9, $p = 0.08$). This increase in visited locations implies that mobility changes are not a result from trivial increases in everyday activity, but also a result from individuals exploring familiar regions more thoroughly (i.e., hypothesis (ii)).

We next examine potential changes in the spatial distribution of mobility by analysing isotropy ratios σ_y/σ_x (see Methods 3.5), i.e., uniformity of mobility. We cluster users by their ratio before the game, at intervals of 0.2, from 0.2 to 0.8, and analyze changes during Pokémon GO use. For users with a high anisotropy, we find that Pokémon GO significantly increases their isotropy, i.e., their geographic distribution of mobility becomes more homogeneously distributed, further supporting hypothesis (ii). For users with $\sigma_y/\sigma_x = 0.2$ before the game, we observed `gamers` to have more isotropic trajectories than the `control` group users during gameplay (0.299 and 0.270 respectively, with $p = 0.016$). For all other clusters and periods there was no statistically significant difference between user groups ($p > 0.05$). Analysis of isotropy thus shows that characteristics of mobility largely remain intact, with only individuals with a low isotropy (i.e., high anisotropy) experiencing changes. These correspond to individuals whose mobility is dominated by long hops, whom Pokémon GO can improve the balance of the mobility distribution.

Given the location-based nature of Pokémon GO, increases in mobility could be associated with higher game playing time instead of an actual effect on physical mobility. To explore this potential bias, we first calculate average session times for Pokémon GO and Clash Royale players from Dataset-A. These are μ_{P_S} = 26 minutes (median: 14.2 minutes, σ: 29.3 minutes) for Pokémon GO, and μ_{P_S} = 28.6 minutes (median: 16.4 minutes, σ: 35 minutes) for Clash Royale, respectively. The usage patterns thus are similar to other non-location-based games. Similar patterns can be observed with the number of days users continue to play Pokémon GO. Specifically, on average, we observe Pokémon GO gamers to play for 99 days (median: 52.8 days, σ: 119.5 days), 21.5 total gaming sessions (median: 7 sessions, σ: 45.9 sessions). For Clash Royale, gamers play on average for 95 days (median: 35 days, σ: 135 days), 16.1 sessions (median: 4 sessions, σ: 40.65 sessions). For both the session times and the playing time we can observe significant differences between mean and median values, suggesting the distributions are heavily skewed. As shown in Sect. 3.7, the session times are also similar across different countries.

4.3 Pokémon GO increases the likelihood of short hops

We next analyze the mobility distribution before and after Pokémon GO to understand how the changes affect the underlying mobility model. We use Dataset.B to analyze the distribution of placements for both the `gamers` and the `control` group. For both groups, mobility is consistent with a truncated Lévy-flight model, but the ratio between short

and long displacements changes between the two groups. Specifically, the distribution of displacements follows a truncated power law combined with an exponential (see Methods). Compared to baseline values shown in Table 3, the value of long-tailed parameter α changes to significantly higher values for `gamers` ($\alpha_g = 0.35 \pm 0.02$) than for `control` ($\alpha_c = 0.33 \pm 0.02$), i.e., `gamers` exhibit an increase in the probability of short hops and a decrease in the probability of long hops compared to `control`.

4.4 Better power management by the app led to greater effects on mobility

Pokémon GO was chosen as representative example of smartphone applications that can promote physical activity, and other applications could have similar effects on mobility. Findings in literature appear mixed on this aspect, showing that many applications only have a temporary effect on mobility [11]. This contrasts with studies on Pokémon GO, which have almost consistently reported sustained change [1–3]. To better understand this discrepancy, we analyzed how the increase in mobility differs across different versions of Pokémon GO, and correlated our findings against technical change-logs (see Sect. 3.9). We find mobility increases to be significant only from Pokémon GO version 0.33 onward, which introduced significant battery saving strategies. Specifically, the impact on daily mobility when starting to play these initial versions of the game is statistically significant only with highly-engaged users (i.e., players of group C, >90 days), whereas all users (i.e., groups A, B and C) starting Pokémon GO with a later version significantly change their mobility while playing the game. Early reports of Pokémon GO linked the app with high battery drain [32], which in turn has been linked with high attrition [33]. Our results suggest that application design mediates mobility changes and that the discrepancy in findings across different applications may be a result from differences in application design and user-friendliness.

5 Discussions

Lack of physical activity has been tightly associated with several health problems [34–36], and ranks high amongst risk factors for premature death as well as disability [37]. Understanding of how location-based games can alter users' mobility can have a significant impact on future policies aimed at incentivizing physical activity [36]. Indeed, physical activity can have health effects comparable to those brought by medications [38], making our findings relevant for physicians and public health.

Mobility laws are of paramount importance for disease transmission modeling as mobility results in opportunities to meet other people and hence creates possible situations where diseases can be transmitted [39–41]. Our results showed that location-based games as an exogenous factor result in higher number of short hops and hence can cause higher local transmission rates [42]. Reversely, while our study focuses on the *increase* of mobility through location-based games, a *reduction* of mobility could be also achieved with the appropriate game elements [15]. Such idea could be implemented in order to curb the spread of infectious diseases by gamifying the adoption of curfew measures as well as social distancing. Physical activity could be retained by, e.g., rewarding users for increased step count or physical activity (e.g., measured by heart rate sensors in a smartwatch) while requiring them to stay within a small geo-fenced area.

Our findings are also interesting to urban scientists and policy makers. The analysis showed how exogenous factors can result in increased exploration of the local region,

which in turn is essential for understanding districts and their dynamics. Our results also offer opportunities for transport and city planning through more detailed mobility models and mechanisms that can be used to shape mobility. Indeed, while extensive literature exists on the utilization of current urban spaces (e.g., [43, 44]), more investigation is needed on how to change such patterns and our work offers an important first step in this direction. For example, a possible example is the use of a location-based game to shape the use of public spaces by driving pedestrians away from congested areas or better planning of public transport.

Pokémon GO had and has a large, worldwide user base, which lends itself well to studying mobility. Applications with smaller and more localized user bases could be used to study the population of a city, speakers of a particular language, a specific socioeconomic group [15], or a geographical area where the app is popular [45]. As these types of applications target specific groups, the characteristics of mobility may differ [46]. Our results shed light on the characteristics and laws that govern the changes when they do occur, showing that re-parameterization of existing models is sufficient to account for the changes in mobility.

Our results corroborate with strong evidence the link between location-based online games and changes in human mobility found in existing research [1, 8, 14], even though our study was limited to a single, albeit exceptional, location-based online game. Other location-based applications, including location-based recommender systems and other location-based games, are likely to have similar effects—provided that they can induce a positive change in the first place. Previous studies on other location-based games have not always shown an effect on mobility, which can be due to lack of suitably engaging content, small user-base, or technical issues. Indeed, our analysis also showed that early versions of Pokémon GO failed to increase mobility and only later versions that improved the end-user experience were successful in motivating people to increase their mobility. Further research is needed to better understand the factors that mediate possible increases in mobility.

6 Conclusion

In this paper, we have studied the mobility laws governing location-based gaming, an important exogenous factor affecting variations in personal mobility across individuals. We analyzed measurements collected through two different means, a location-based social network (Twitter) and mobile app-logging. Our results show that exposure to location-based gaming can significantly influence an individual's mobility but the characteristics governing mobility remain consistent with a truncated Lévy-flight model. The main difference in mobility is an increased degree of short hops, evidenced from a more homogeneous isotropy ratio, but unaffected radius of gyration. We showed that mobility changes are explainable by a higher degree of exploration of previously visited regions, instead of a consistent change in mobility patterns. Our results improve our collective understanding of human mobility, demonstrating how exogenous factors can help to explain inter-individual variations, and showing how these variations can be modeled using prevalent mobility models with adjustments to their parameters. Specifically, variations in an individual's mobility can be captured using personal-level models that account for the individual's exposure to different factors. Taken together, our results corroborate the effect of location-based online games of changes in human mobility found in existing research

[1, 8, 14], while offering novel insights into the laws governing the characteristics of these changes.

Beyond improving our collective understanding of mobility, the results provide insights into mobility characteristics of location-based smartphone applications and provide suggestions on how to improve their potential in promoting physical activity. For example, the game mechanics of Pokémon GO have been designed around so-called (Poke)stops, which are important locations around which the game activity is centered. Previous research has shown that these stops are not uniformly distributed, but cover different regions of cities [15], with a strong bias towards densely populated urban areas where most of mobility already took place before the game. Our results showed that exploration largely takes place in close proximity of previously visited places, suggesting that stops or other focal areas near familiar regions have the highest likelihood of attracting the user.

Acknowledgements
Not applicable.

Funding
Open Access funding enabled and organized by Projekt DEAL.

Abbreviations
GPS, Global Positioning System; DBSCAN, Density-Based Spatial Clustering of Applications with Noise; REST, Representational State Transfer; API, Application Programming Interface; OS, Operating System.

Availability of data and materials
The anonymized version of the Carat data that support the findings of this study are available at https://www.cs.helsinki.fi/group/carat/mobility-games-data/.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
EL was responsible for the Carat dataset, LT collected the remaining datasets. LT, EL and PN conceived the research idea and co-wrote the manuscript with input from all authors. LT and EL analyzed the data, and PN, JO, ST and AD supervised the project. All authors read and approved the final manuscript.

Author details
[1]Techinical University of Munich, Boltzmannstrasse 3, 85748 Garching bei München, Germany. [2]University of Helsinki, 00014 Helsinki, Finland. [3]TU Delft, 2600GA Delft, The Netherlands.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 6 October 2020 Accepted: 3 February 2021 Published online: 15 February 2021

References

1. Althoff T, White RW, Horvitz E (2016) Influence of Pokémon Go on physical activity: study and implications. J Med Internet Res 18(12):315
2. Xu H, Xian Y, Xu H, Liang L, Hernandez AF, Wang TY, Peterson ED (2017) Does Pokémon Go help players be more active? An evaluation of Pokémon Go and physical activity. Circulation 135(suppl_1):02–02
3. Khamzina M, Parab KV, An R, Bullard T, Grigsby-Toussaint DS (2020) Impact of Pokémon Go on physical activity: a systematic review and meta-analysis. Am J Prev Med 58(2):270–282
4. Lear SA, Hu W, Rangarajan S, Gasevic D, Leong D, Iqbal R, Casanova A, Swaminathan S, Anjana RM, Kumar R et al (2017) The effect of physical activity on mortality and cardiovascular disease in 130 000 people from 17 high-income, middle-income, and low-income countries: the pure study. Lancet 390(10113):2643–2654
5. Hu K, Der Riemersma-Van LRF, Patxot M, Li P, Shea SA, Scheer FA, Van Someren EJ (2016) Progression of dementia assessed by temporal correlations of physical activity: results from a 3.5-year, longitudinal randomized controlled trial. Sci Rep 6(1):1–10
6. Althoff T, Jindal P, Leskovec J (2017) Online actions with offline impact. In: Proceedings of the ACM WSDM, pp 537–546. arXiv:1612.03053
7. Wong FY (2017) Influence of Pokémon Go on physical activity levels of university players: a cross-sectional study. Int J Health Geogr 16(1):17
8. Gal R, May AM, van Overmeeren EJ, Simons M, Monninkhof EM (2018) The effect of physical activity interventions comprising wearables and smartphone applications on physical activity: a systematic review and meta-analysis. Sports Med 4(1):42

9. Cho E, Myers SA, Leskovec J (2011) Friendship and mobility: user movement in location-based social networks. In: Proceedings of the ACM SIGKDD, pp 1082–1090
10. Shameli A, Althoff T, Saberi A, Leskovec J (2017) How gamification affects physical activity: large-scale analysis of walking challenges in a mobile application. In: Proceedings of the international conference on World Wide Web, pp 455–463. https://doi.org/10.1145/3041021.3054172
11. Romeo A, Edney S, Plotnikoff R, Curtis R, Ryan J, Sanders I, Crozier A, Maher C (2019) Can smartphone apps increase physical activity? Systematic review and meta-analysis. J Med Internet Res 21(3):12053
12. Lin JJ, Mamykina L, Lindtner S, Delajoux G, Strub HB (2006) Fish'n'steps: encouraging physical activity with an interactive computer game. In: UbiComp. Springer, Berlin, pp 261–278
13. Munson SA, Consolvo S (2012) Exploring goal-setting, rewards, self-monitoring, and sharing to motivate physical activity. In: International conference on pervasive computing technologies for healthcare and workshops. IEEE, New York, pp 25–32
14. Bravata DM, Smith-Spangler C, Sundaram V, Gienger AL, Lin N, Lewis R, Stave CD, Olkin I, Sirard JR (2007) Using pedometers to increase physical activity and improve health: a systematic review. JAMA 298(19):2296–2304
15. Colley A, Thebault-Spieker J, Lin AY, Degraen D, Häkkilä J, Kuehl K, Nisi V, Nunes NJ, Wenig N, Wenig D, Hecht B, Schöning J The geography of Pokémon GO: beneficial and problematic effects on places and movement. In: CHI 2017 (2017)
16. Oliner AJ, Iyer AP, Stoica I, Lagerspetz E, Tarkoma S (2013) Carat: collaborative energy diagnosis for mobile devices. In: Proceedings of the 11th ACM conference on embedded networked sensor systems, 10–11014
17. Peltonen E, Lagerspetz E, Nurmi P, Tarkoma S (2015) Energy modeling of system settings: a crowdsourced approach. In: 2015 IEEE international conference on pervasive computing and communications, PerCom 2015, pp 37–45
18. Varol O, Ferrara E, Davis CA, Menczer F, Flammini A (2017) Online human–bot Interactions: detection. Estimation, and characterization. arXiv:1703.03107
19. Rasche P, Schlomann A, Mertens A (2017) Who is still playing Pokémon Go? A web-based survey. JMIR Serious Games 5(2):7
20. Ester M, Kriegel H-P, Sander J, Xu X et al (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd, vol 96, pp 226–231
21. Ashbrook D, Starner T (2003) Using gps to learn significant locations and predict movement across multiple users. Pers Ubiquitous Comput 7(5):275–286
22. González MC, Hidalgo CA, Barabási A-L (2008) Understanding individual human mobility patterns. Nature 453:1–12. arXiv:0806.1256
23. Zhao K, Musolesi M, Hui P, Rao W, Tarkoma S (2015) Explaining the power-law distribution of human mobility through transportation modality decomposition. Sci Rep 5(1):1–7
24. Sobkowicz P, Thelwall M, Buckley K, Paltoglou G, Sobkowicz A (2013) Lognormal distributions of user post lengths in Internet discussions—a consequence of the Weber–Fechner law? EPJ Data Sci 2(1):2
25. Jurdak R, Zhao K, Liu J, AbouJaoude M, Cameron M, Newth D (2015) Understanding human mobility from Twitter. PLoS ONE 10(7):0131469
26. Peltonen E, Lagerspetz E, Hamberg J, Mehrotra A, Musolesi M, Nurmi P, Tarkoma S (2018) The hidden image of mobile apps: geographic, demographic, and cultural factors in mobile usage. In: Proceedings of the 20th international conference on human–computer interaction with mobile devices and services. ACM, New York, p 10
27. Leontiadis I, Lima A, Kwak H, Stanojevic R, Wetherall D, Papagiannaki K (2014) From cells to streets: estimating mobile paths with cellular-side data. In: Proceedings of the 10th ACM international on conference on emerging networking experiments and technologies. CoNEXT '14. ACM, New York
28. Kitamura R, Van Der Hoorn T (1987) Regularity and irreversibility of weekly travel behavior. Transportation 14(3):227–251
29. Goulet-Langlois G, Koutsopoulos HN, Zhao Z, Zhao J (2017) Measuring regularity of individual travel patterns. IEEE Trans Intell Transp Syst 19(5):1583–1592
30. Sigg S, Lagerspetz E, Peltonen E, Nurmi P, Tarkoma S (2019) Exploiting usage to predict instantaneous app popularity: trend filters and retention rates. ACM Trans Web 13(2):13
31. Athukorala K, Lagerspetz E, Von Kügelgen M, Jylhä A, Oliner AJ, Tarkoma S, Jacucci G (2014) How carat affects user behavior: implications for mobile battery awareness applications. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, pp 1029–1038
32. Paavilainen J, Korhonen H, Alha K, Stenros J, Koskinen E, Mayra F (2017) The Pokémon Go experience: a location-based augmented reality mobile game goes mainstream. In: Proceedings of the 2017 CHI conference on human factors in computing systems. ACM, New York, pp 2493–2498
33. Zuniga A, Flores H, Lagerspetz E, Tarkoma S, Manner J, Hui P, Nurmi P (2019) Tortoise or hare? Quantifying the effects of performance on mobile app retention. In: International World Wide Web conference on World Wide Web (WWW 2019). International World Wide Web Conferences Steering Committee
34. Blair SN (2009) Physical inactivity: the biggest public health problem of the 21st century. Br J Sports Med 43(1):1–2
35. Warburton DE, Nicol CW, Bredin SS (2006) Health benefits of physical activity: the evidence. CMAJ, Can Med Assoc J 174(6):801–809
36. Ding D, Lawson KD, Kolbe-Alexander TL, Finkelstein EA, Katzmarzyk PT, van Mechelen W, Pratt M (2016) The economic burden of physical inactivity: a global analysis of major non-communicable diseases. Lancet 388(10051):1311–1324. https://doi.org/10.1016/S0140-6736(16)30383-X
37. Murray CJ, Abraham J, Ali MK, Alvarado M, Atkinson C, Baddour LM, Bartels DH, Benjamin EJ, Bhalla K, Birbeck G et al (2013) The state of us health, 1990–2010: burden of diseases, injuries, and risk factors. JAMA 310(6):591–606
38. Naci H, Ioannidis JP (2013) Comparative effectiveness of exercise and drug interventions on mortality outcomes: metaepidemiological study. BMJ, Br Med J 347:5577
39. Kraemer M, Golding N, Bisanzio D, Bhatt S, Pigott D, Ray S, Brady O, Brownstein J, Faria N, Cummings D et al (2019) Utilizing general human movement models to predict the spread of emerging infectious diseases in resource poor settings. Sci Rep 9(1):1–11
40. Colizza V, Barrat A, Barthélemy M, Vespignani A (2006) The role of the airline transportation network in the prediction and predictability of global epidemics. Proc Natl Acad Sci USA 103(7):2015–2020

41. Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M, Merler S, Pastore y Piontti A, Mu K, Rossi L, Sun K et al (2020) The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science 368:395–400
42. Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, Shaman J (2020) Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science 368:489–493
43. Liu Y, Kang C, Gao S, Xiao Y, Tian Y (2012) Understanding intra-urban trip patterns from taxi trajectory data. J Geogr Syst 14(4):463–483
44. Horner MW, Downs JA (2014) Integrating people and place: a density-based measure for assessing accessibility to opportunities. J Transp Land Use 7(2):23–40
45. Schechtner K, Hanson M (2017) Shared mobility in Asian megacities: the rise of the apps. In: Disrupting mobility. Springer, Berlin, pp 77–88
46. Lima A, De Domenico M, Pejovic V, Musolesi M (2015) Disease containment strategies based on mobility and information dissemination. Sci Rep 5:10650

Submit your manuscript to a SpringerOpen journal and benefit from:

► Convenient online submission
► Rigorous peer review
► Open access: articles freely available online
► High visibility within the field
► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com

Publication 5

©2022 Elsevier, reprinted with permission from:

Tonetto, L., Adikari, M., Mohan, N., Ding, A. Y., and Ott, J. (2022). Contact duration: Intricacies of human mobility. In Online Social Networks and Media (28, 100196).

RightsLink

Contact duration: Intricacies of human mobility

Author: Leonardo Tonetto,Malintha Adikari,Nitinder Mohan,Aaron Yi Ding,Jörg Ott

Publication: Online Social Networks and Media

Publisher: Elsevier

Date: March 2022

© 2022 The Authors. Published by Elsevier B.V.

Creative Commons

This is an open access article distributed under the terms of the Creative Commons CC-BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.

To request permission for a type of use not listed, please contact Elsevier Global Rights Department.

Are you the author of this Elsevier journal article?

© 2024 Copyright - All Rights Reserved | Copyright Clearance Center, Inc. | Privacy statement | Data Security and Privacy | For California Residents | Terms and ConditionsComments? We would like to hear from you. E-mail us at customercare@copyright.com

Title: Contact duration: Intricacies of human mobility
Authors: **Leonardo Tonetto** (TUM), Malintha Adikari (TUM), Nitinder Mohan (TUM), Aaron Yi Ding (TU Delft), Jörg Ott (TUM)
Journal: Online Social Networks and Media (28)
Publishing date: December 19, 2021
Reference: [6]

Publication Summary

In this article, we investigate the connection between human mobility and contact duration, focusing on their implications for studying infectious diseases and content forwarding. The study utilizes data from a mobile social network consisting of 71 individuals, collected over a two-month period in 2018. The data include GPS and Bluetooth logs, which are augmented with location APIs to provide more detailed information about users' mobility and contact patterns.

The analysis of stops (or stays) reveals that their durations can be modeled using different probability distributions. Time-unbounded stops, such as bars or restaurants, follow a log-normal distribution, while time-bounded stops, such as offices or hotels, follow a power-law distribution. These findings are consistent with previous studies that observed similar distributions in web-content viewing time, where time-free content follows power-law distributions and time-correlated content follows log-normal distributions.

We also investigate inter-personal contact duration and find that it adheres to a log-normal distribution. Based on this observation, we propose a model to estimate contact durations as a function of the overall duration of stays. Additionally, we analyze contact duration during trips and find that it follows a Weibull distribution, with parameters depending on the trip length.

The results of this study have implications for modeling information or epidemic spreading and can inform the design of network protocols and policy decisions. By better understanding human mobility and contact patterns, it becomes possible to develop more accurate models for studying infectious diseases and guiding measures to control their spread.

Contribution

I came up with the idea and designed the experiments with the supervision from Jörg Ott. Malintha Adikari analyzed the data. I wrote the text with support from Malintha. All authors reviewed the text.

Online Social Networks and Media 28 (2022) 100196

Contents lists available at ScienceDirect

Online Social Networks and Media

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

Contact duration: Intricacies of human mobility

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

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

ARTICLE INFO

Keywords:
COVID-19
Human mobility
Bluetooth sensing
Opportunistic forwarding

ABSTRACT

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

1. Introduction

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

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

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

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

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

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

https://doi.org/10.1016/j.osnem.2021.100196
Received 30 June 2021; Received in revised form 10 November 2021; Accepted 19 December 2021
Available online 19 January 2022
2468-6964/© 2022 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

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

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

2. Related work

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

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

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

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

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

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

3. Background

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

Basic definitions

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

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

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

Empirical distribution functions

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

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

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

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

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

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

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

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

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

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

4. Data collection

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

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

4.1. Location data

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

4.2. Bluetooth contact data

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

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

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

Table 1
Summary of the data set used.

Users	Stops	Encounters	Trips
71	19317	12432	18438

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

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

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

4.3. Supporting set - Geolife

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

5. Stops

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

5.1. Detection of stops

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

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

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

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

5.2. Stops enrichment

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

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

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

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

5.3. Stops duration

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

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

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

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

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

5.4. Contacts characterization at stops

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

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

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

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

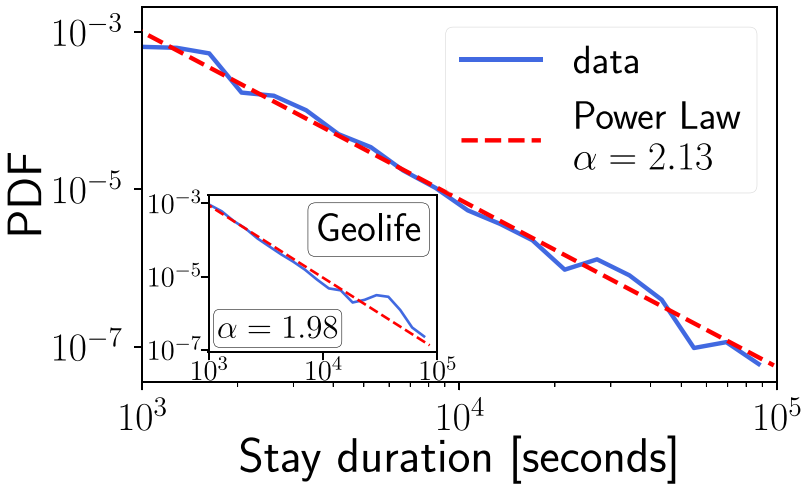


Fig. 1. Overall stop duration follows a **power-law** distribution.

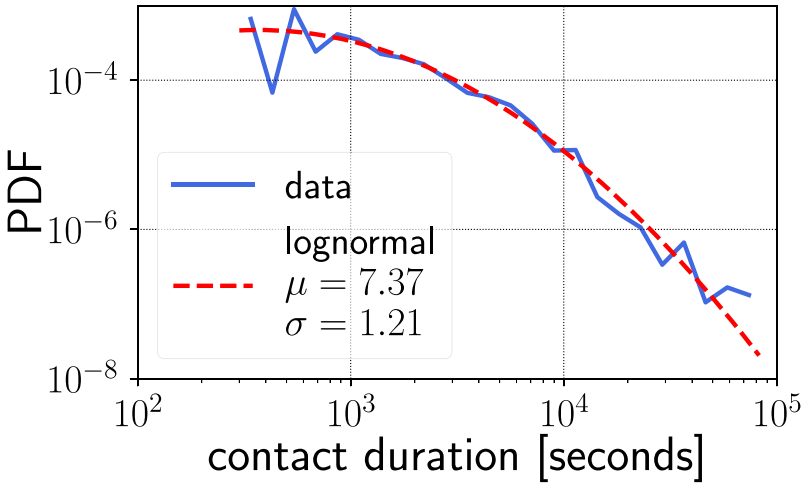


Fig. 2. Contact duration at stops follows a **log-normal** distribution.

5.5. Model of contacts during stops

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

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

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

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

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

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

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

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

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

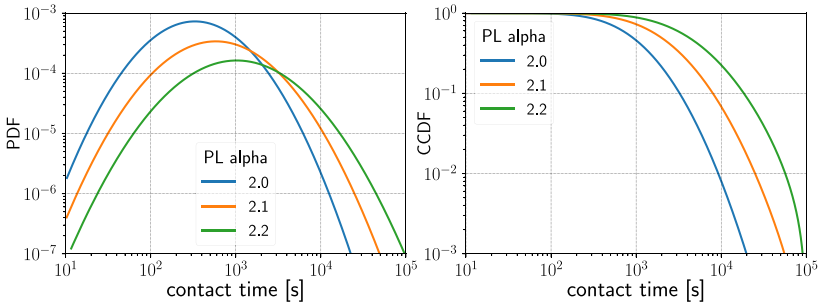


Fig. 3. Distribution of modeled contact duration for different values of the stay duration parameter (*alpha*). Larger values of *alpha* for stay duration indicate *higher* probability for shorter stays, leading to an increase in the probability of long-term contacts as short-term meets become less often.

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

6. Trips

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

6.1. Detection of trips

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

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

6.2. Trip duration and total length

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

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

6.3. Model of contacts during trips

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

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

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

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

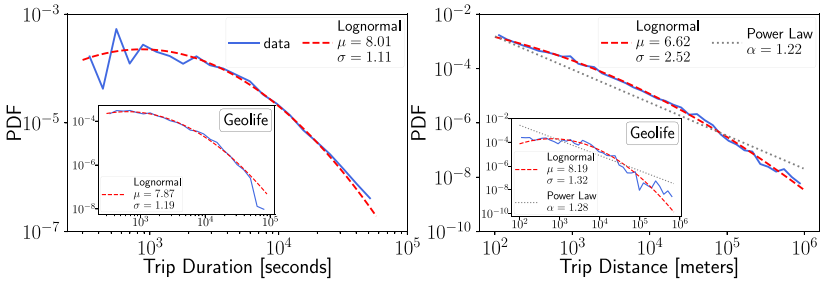


Fig. 4. Both trip time duration and length are best modeled by a log-normal.

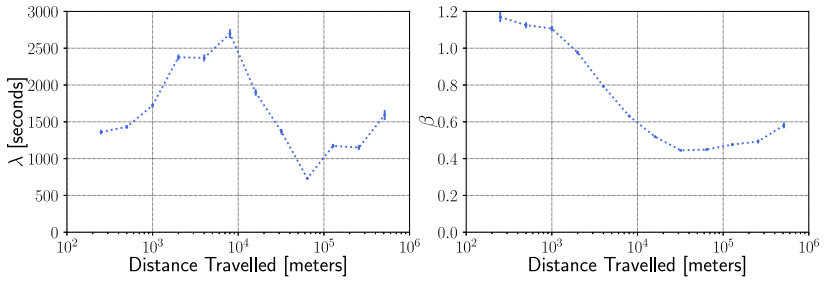


Fig. 5. Variation of the Weibull parameters as a function of distance traveled.

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

7. SIR model and implications

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

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

8. Limitations and discussions

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

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

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

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

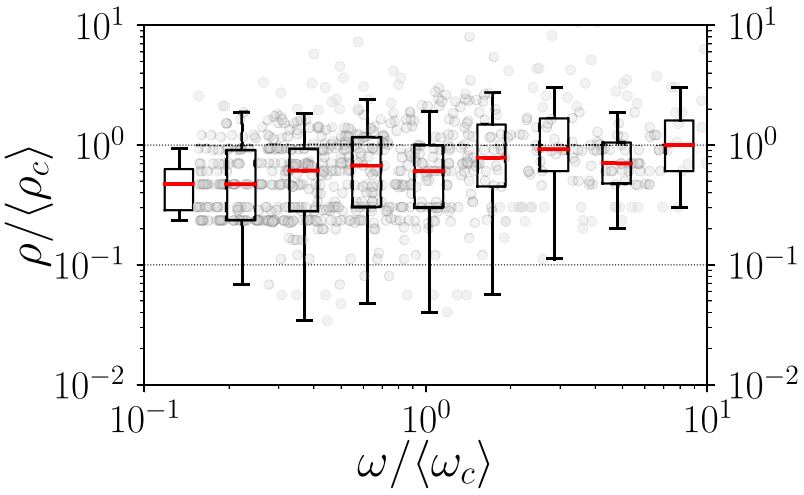


Fig. 6. Number of contacts increases with longer stays (Spearman correlation = 0.3, p-value <0.01). (Red) Lines inside boxes represent median values.

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

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

9. Conclusion

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

10. Ethical considerations

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

CRediT authorship contribution statement

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

Declaration of competing interest

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

Acknowledgments

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

References

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

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

Publication 6

©2022 IEEE, reprinted with permission from:

L. Tonetto, A. Carrara, A. Y. Ding and J. Ott, "Where Is My Tag? Unveiling Alternative Uses of the Apple FindMy Service," 2022 IEEE 23rd International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), Belfast, United Kingdom, 2022, pp. 396-405, doi: 10.1109/WoWMoM54355.2022.00059.

RightsLink

Sign in/Register

Where Is My Tag? Unveiling Alternative Uses of the Apple FindMy Service

IEEE Requesting permission to reuse content from an IEEE publication

Conference Proceedings:
2022 IEEE 23rd International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)

Author: Leonardo Tonetto

Publisher: IEEE

Date: June 2022

Copyright © 2022, IEEE

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK CLOSE WINDOW

© 2024 Copyright - All Rights Reserved | Copyright Clearance Center, Inc. | Privacy statement | Data Security and Privacy | For California Residents | Terms and ConditionsComments? We would like to hear from you. E-mail us at customercare@copyright.com

Title: Where Is My Tag? Unveiling Alternative Uses of the Apple FindMy Service
Authors: **Leonardo Tonetto** (TUM), Andrea Carrara (TUM), Aaron Yi Ding (TU Delft), Jörg Ott (TUM)
Venue: 23rd International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), Belfast, Northern Ireland
Publishing date: June 14-17, 2022
Reference: [1]

Publication Summary

In this paper, we evaluate Bluetooth trackers (*i.e.*, tags) on Apple's FindMy network for crowd monitoring, deliberate tracking, and covert communication. We present approaches to crowd size estimation using a single tag, crowd flow analysis using multiple tags, deliberate tracking with examples of remote destination inference and path reconstruction, and covert communication using a single tag. We also present threat models, proof of concept (PoC) descriptions, results, and mitigation strategies for the demonstrated attacks.

On crowd monitoring, we evaluate the effectiveness of tags in estimating crowd size by comparing the results with image recognition. We analyze the correlation between the number of identified finders using smart tags and the number of people identified through image recognition.

On crowd flow, we evaluate the use of tags for studying the flow of a crowd. We compare our method with the use of Wi-Fi management frames. We describe the use of Wi-Fi management frames and their tracking capabilities by analyzing the distribution of time intervals between two vantage points using tags and Wi-Fi measurements.

On deliberate tracking, we present proof of concept evaluations and mitigation strategies for information leakage through the Apple FindMy service. We demonstrate how timing attacks can be used to track a victim's device and infer their destinations and paths. We discuss the threat models, and results of a proof-of-concept attack, suggesting mitigation strategies such as reducing the granularity of timestamps and randomizing the uploading of reports.

On covert communication, we present TagComm, a protocol built on the FindMy service, which uses nearby iPhones to encode and transport information through the sensing of tags IDs. The protocol utilizes permutations of tag IDs to encode the information. Additional header bits and a parity bit are included to ensure transmission integrity. The experiments conducted to test TagComm demonstrate successful transmission of random words with varying parameters. The error rate was affected by frame duration and BLE advertising intervals, with larger intervals increasing the error rate.

Overall, we discuss the utilization of Bluetooth trackers (tags) beyond their original purpose of tracking lost devices. We conduct experiments to demonstrate the privacy risks associated with the FindMy service, which can disclose sensitive location information. Furthermore, we highlight the possibility of reconstructing a person's path and visits using this information. We also introduce a proof-of-concept out-of-band

communication channel using crafted tags, enabling the encoding and transmission of arbitrary information securely. Finally, we emphasize the need to raise awareness about these possibilities and discuss potential uses, including side-channel communication that could leak sensitive information from a compromised system.

Contribution

I came up with the idea with the supervision from Jörg Ott. I designed the experiments and collected the data. Andrea Carrara executed the image recognition step of the analysis. I wrote the manuscript. All authors reviewed the text.

Where Is My Tag? Unveiling Alternative Uses of the Apple FindMy Service

Leonardo Tonetto, Andrea Carrara
Department of Computer Science
Technical University of Munich
Munich, Germany
tonetto@in.tum.de, andrea.carrara@tum.de

Aaron Yi Ding
Dept. of Engineering Systems and Services
Delft University of Technology
Delft, Netherlands
aaron.ding@tudelft.nl

Jörg Ott
Department of Computer Science
Technical University of Munich
Munich, Germany
ott@in.tum.de

***Abstract*—Bluetooth trackers, or tags, have quickly become ubiquitous and widely supported by multiple vendors. Beyond their original design of finding lost objects, these devices have the ability to extend the capabilities of current wireless smart devices. Since its launch in 2019, Apple's FindMy enables any devices from their brand to be easily tracked by more than 1 billion active iPhones and iPads on the market. While convenient, these systems may even serve further uses, including as a result of this work, crowd sensing and a side channel for mobile communication. But they also raise privacy concerns for their users. In this paper, we demonstrate how Apple FindMy can be used as a privacy-friendly tool for crowd monitoring, and how it may inadvertently leak information on a person's location in case of deliberate tracking. Additionally, we design and evaluate a proof of concept protocol, using the Apple FindMy and a crafted tag using a simple microcontroller. We show how such system could be used to transmit information at very low bit rates, while the devices transporting the information remain unaware of this covert channel, yielding an *out of band* communication channel.**

***Index Terms*—sensing, location privacy, crowd monitoring, mobile communication, covert exfiltration**

I. Introduction

Bluetooth trackers, or tags, have become ubiquitous, being primarily used to track and find lost objects. This growing pervasiveness allowed manufacturers to create a network of tracker owners to *anonymously* report about any nearby tag, such as Tile and Apple FindMy. This crowdsourced reporting provides obvious primary benefits for its users, but also enables alternative uses for which it was not originally designed.

In this paper, we explore two such alternative uses as the main objective of our work: (1) a fully anonymous crowd sensing system and (2) a covert communication channel built on top of Apple's FindMy system: As an auxiliary finding, our work revealed potential privacy issues that could affect billions of Apple devices [1], for which the only current mitigation is disabling Bluetooth or opting-out of the FindMy service.

Crowd Sensing: Sensing and monitoring different aspects of a (large) crowd may serve numerous purposes, such as steering people flows to safety under pressing conditions. However, automated methods for crowd monitoring, such as image-based tracking, may raise privacy concerns [2]. Individuals are bothered by sensors capturing any form of personal identifiable information (PII) that, if stored permanently, may require explicit consent from those being monitored.

Current solutions to this privacy problem in crowd monitoring may rely on computing all relevant metrics at the edge [2]. However, these systems still handle PII and those being monitored simply have to trust their personal data are being dealt with appropriately. Therefore, a reliable source of crowd data while guaranteeing the privacy of its subjects is still a relevant open problem that we explore with this paper.

We use handcrafted Apple tags enabled by the reverse-engineering work of Heinrich *et al.* [1], in which single board computers and microcontrollers can be used as tags. Apple allows the owner of tags to download all location reports within a week. Through a series of comprehensive analyses, we demonstrate the capability of such system using trackers, or sensory-tags, for coarse crowd monitoring, including determining counts/density and flows.

Covert Data Channel: We also demonstrate how such tracker systems could be used to create a side channel for communication, while sending information *silently* through nearby mobile devices. Our proof-of-concept enables out-of-band communication at low bit-rate, without awareness of those partially carrying the information.

Deliberate Tracking: We explore potential privacy risks associated with Apple FindMy as a side effect of its sensing capabilities. The threats we reveal concern the possibility of exposing location information of a victim from the timestamps contained in each location report. We demonstrate their feasibility through proof of concept examples and discuss possible mitigation approaches to these threats.

Our work exposes and discusses alternative usages for an established secure system, including potential malicious ones. We present an evaluation of the Apple FindMy network and its main properties that are pertinent to the proposed solutions. Furthermore, we design and test a simple protocol, with basic characteristics to ensure a successful transmission of data with our system. Finally, we discuss the implications of this work, along with possible mitigation strategies for users and developers of similar systems. *We reported all uncovered issues to Apple several months prior to the submission of this manuscript.*

Our contributions: (1) We thoroughly analyse the timing and conditions in which location reports are sent for a lost smart tag in the Apple finder network (§ IV). (2) We demonstrate the feasibility and accuracy of using such anonymous location reports for sensing two different aspects of a crowd, namely flow and size/density, from a series of real-world measurements compared to state of the art solutions (§ VI). (3) We reveal the feasibility of timing attacks, using the Apple finder network that could reveal information about a person's whereabouts without their consent (§ VII). (4) We show such tags can be used to transmit information through a side-channel, and we name it TagComm (§ VIII). (5) Complementing the original work by Heinrich *et al.* [1], we provide open source code that enables other devices to be used as sensory-tags as well as be used for covert communication: `macOS` devices and the Amazon Echo [3], [4]. (6) We discuss the privacy and ethical implications of our work (§ IX).

II. Related Work

1) Apple Ecosystem: Recently, various papers evaluated the security of different Apple services. Analysis by Martin *et al.* of the Handoff services, that enables seamless communication between multiple Apple devices under the same iCloud account, reveals how Apple's proprietary protocol can undermine MAC address randomization and allow the identification of devices belonging to a single user [5]. Furthermore, a study by Stute *et al.* demonstrated how the Apple Wireless Direct Link (AWDL) works, including the reverse engineering of its protocols and Wireshark plugins. These examples support the importance of scrutinizing such proprietary systems that may affect users of billions of devices worldwide [6].

2) Bluetooth LE trackers: A recent study by Weller *et al.* evaluated different Bluetooth trackers and their cloud services [7]. Their study revealed a series of security issues, including privacy risks with all products tested, although it did not include Apple's FindMy as no commercial product was available at the time. Focusing exclusively on the Apple service, Heinrich *et al.* dissected how FindMy components work [1]. Their study reverse engineered the protocol used by lost devices, finders and how owners can retrieve available location reports for their tags. Their open source code was used as the foundation for our present paper.

3) Security and Privacy: Security and privacy literature has a vast number of systems that exploit different vectors to covertly exfiltrate data from systems (*e.g.* [8]). Various systems have used keyboard or HDD indicator LEDs [9], as well as inaudible (or indistinguishable) sound from speakers [10], fans or HDDs [11]. In this paper, the proof-of-concept we present enables a covert channel, through which any information can transmitted at low bit rates. To foster further research, we extend [1] by `macOS` support that runs without `root` privilege [3], [4].

4) Crowd Monitoring: Assessing and understanding large crowds has been studied with a myriad of sensors and methods, but not without its privacy implications [2]. However, several challenges are still open when it comes to scalability and integration of multiple systems towards a common decision support [12]. The COVID-19 pandemic has stimulated crowd monitoring research, for example, ensuring social distancing [13] as a valuable approach to reduce infections [14].

5) Our Work: In this paper, we further analyse the Apple FindMy service, and we present two proof-of-concept systems, one which allows coarse crowd monitoring, and other that allows side-channel communication as well as their potential risks for users' privacy. Note that, while [1] reverse engineered the client-side managing of tags, we extend our understanding of this system while exposing possible security and privacy leaks FindMy users are currently subject to.

III. Background

In this section, we present basic functionality of Apple FindMy as the underlying system for our current work. Furthermore, we present relevant concepts of Bluetooth LE.

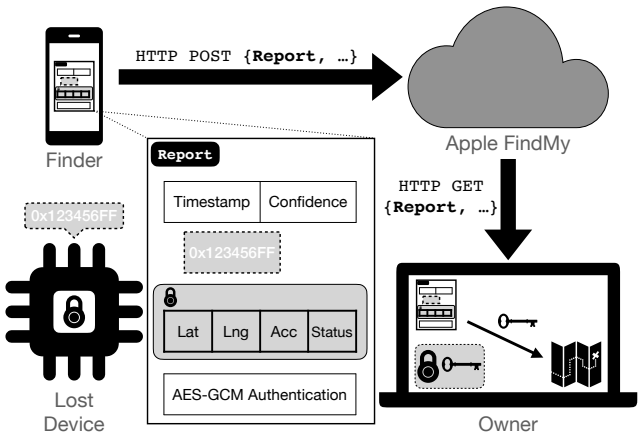


Fig. 1: Delay in sensing and reporting a tag.

A. Apple FindMy Service

This finder network was released in 2019, in which devices that explicitly opt-in are tracked through anonymous crowd-sourced location reports. When devices are marked as lost, their owners receive location reports through their iCloud account and view them, e.g., using the FindMy application [15].

Heinrich *et al.* [1] reverse-engineered this system, allowing a series of Bluetooth Low Energy (BLE) devices to appear as tags inside Apple's FindMy network. Any of such tags will beacon at all times, regardless of the presence of its owner. Furthermore, when marked as *lost* the owner of a tag may then receive any available location report. To get started, an iCloud user, the *owner* of a tag, creates a public-private key pair (e_k, d_k) through a series of API calls, for tracking a device.

1) Beaconing: To enable tracking, a device broadcasts a BLE advertising packet using a specific MAC address that is derived from the above public key e_k. *Finder* devices listen for Apple FindMy beacons and check for each received beacon if the advertised payload and MAC address are a valid "match". A tag derives both payload and MAC address from the public key (e_k) created by its owner (see [1] for details

on their algorithm), so that only valid packets are processed further. *Finder* devices receiving such beacons can then furnish location reports for nearby tags, anonymously to iCloud.

2) Reporting a tag: The reporting process is depicted in Figure 1. A tag broadcasts an appropriate BLE beacon, as described above. A finder device passing by will identify this as a lost device and store a location report, containing (1) the beacon reception time, termed contact (t_c), (2) the confidence about that contact (similar to accuracy in (4)), (3) the public key e_k, (4) the location information, encrypted using e_k and containing geographical coordinates, horizontal accuracy, and status, and (5) an authentication label AES-GCM to validate the report. These reports are uploaded securely (HTTPS POST) after some time to Apple's iCloud, where they are stored until being requested by the tag owner. In addition to these data fields, a *bundle* of reports submitted by a single finder is annotated with the timestamp of when the entire batch was received on the server side (t_r).

3) Reading reports: With the public keys (e_k) of their own tags, users query their iCloud account for available location reports with HTTPS GET requests; each user can then decrypt the location information contained in a report using the corresponding private key (d_k).

B. BLE Advertising and Our Experimental Setup

The BLE standard allows advertising packets of devices to include up to 31 bytes of information. These beacons are broadcast at intervals between 20 ms and 10 s on any of three channels used for advertisements [16]. Their successful reception by a nearby finder device is stochastic: the transmission (TX) power for the advertising packets may influence proper reception and may the distance between finder and tag and the environmental conditions (e.g., radio interference). Also, both finder and tag continuously switch channels and would need to use the same one when a packet is sent.

Our Setup: Given these constraints, and to better understand the conditions in which locations of devices are reported, we carry out a series of experiments to build a thorough understanding of how sensing and reporting work in the Apple FindMy network. We use a series of ESP32 microcontrollers as tags for our experiments. These low-power devices provide programmable BLE support through an API which allows full control of its Bluetooth controller, including TX power and advertising interval, which are often not accessible on other platforms. These experiments allow us to draw observations which set the foundation to the side-channel communication we discuss on the following sections.

IV. FindMy System Characterization

In this section, we present the results of a series of experiments we conducted in controlled environments as well as in the wild. We first present our findings on the behavior of Apple devices when reporting smart tags to the FindMy network. Next, based on observations drawn from our aforementioned findings, we present results from crowd monitoring measurements compared to state of the art alternatives, as well as a possible side-channel attack.

A. Uploading of reports is determined by device settings

As discussed in Section III, finder devices often bundle a series of reports before uploading them to iCloud. To better characterize this behavior, we analyzed the traffic between iCloud and two jailbroken[1] iPhones (7 and 8, on iOS14.6) using an HTTP proxy. With a Bluetooth tracker, continuously beaconing, placed next to these phones for intervals of 72 hours, we tested how different settings influenced the uploading intervals. We present the distribution of these intervals in Figure 2 for various settings, with a clear distinction between being on Wi-Fi (median ∼15 minutes) or Cellular (median ∼3 hours), on power supply or battery. Additionally, with Low Data Mode enabled, the phones uploaded reports less often (median ∼36 minutes), whereas other modes did not affect reporting significantly. Therefore, the phone settings explain differences between the contact time t_c reported and received time t_r, *when a bundle of reports is sent*.

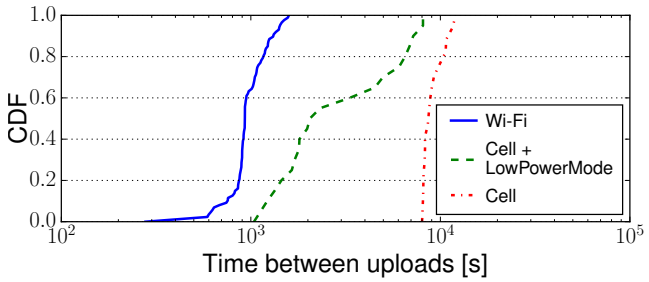


Fig. 2: Delay in sensing and reporting a SmartTag.

B. Over 50% of reports are uploaded within 15 minutes

We ran a set of measurements in the wild, at a large public space. These observations lasted for a total 24 hours, and were conducted on various days from July to September 2021. From these data, we computed the delay between sensing a tag (t_c) and uploading the reports to iCloud (t_r), for which the distribution is depicted in Figure 3. This delay shows a strong mode around 15 minutes, a median of 13.15 minutes, and has 95% of its values between 6 seconds and 8 hours (shaded area).

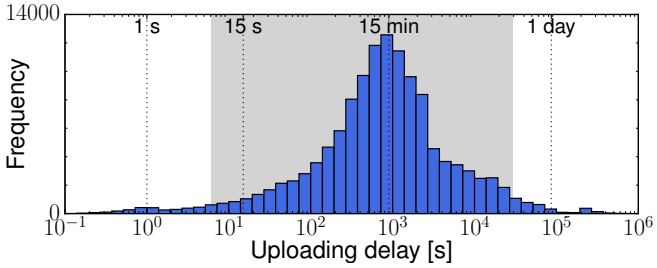


Fig. 3: Distribution of the delay in sensing and uploading.

[1]Required as iCloud HTTPS communication requires certificate pinning.

C. Uploading time uniquely identifies a finder

As reports are bundled, their receiving time t_r is appended on the server side with the precision of milliseconds. This, in turn, allows us to *uniquely* identify a finder for a series of reports as those will contain the same t_r with several decimal points of precision. As discussed above, the uploading of reports may be done hours apart from their actual contact time. It is important to note that each upload may contain up to 255 reports and up to 4 per tag. That is, if a nearby finder is connected to Wi-Fi, only up to 4 location reports will be uploaded every 15 minutes.

D. Short advertising intervals lead to no reports

During our measurements in the wild using short BLE advertising intervals (*e.g.*, 20 ms), we observed that the FindMy network discards *all* location reports for a tag, possibly due to uploads happening too frequently. Although we were not able to precisely determine the best limit, the fastest we could advertise without periods of missing data was 1022.5 ms. For that, in all measurements discussed in Section VI we carried tags configured at that BLE advertising interval (as suggested by Apple [17]) and at maximum TX power (+9 dBm).

Takeaway (§IV): FindMy provides limited but valuable information on nearby finders and that depends on the settings of their mobile devices. Moreover, the reports upload may be delayed from 15 minutes to several hours.

V. General Overview

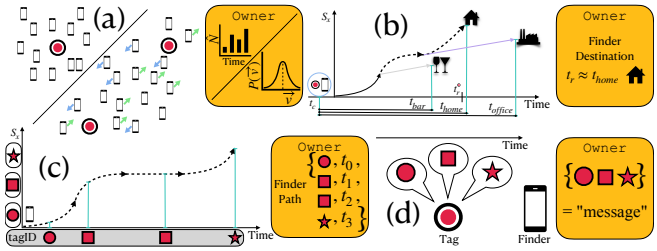


Fig. 4: Overview of alternative uses for the Apple FindMy service. (a) Crowd monitoring. (b) Path reconstruction. (c) Remote destination inference. (d) Covert communication.

Given the observations drawn from the characterization of FindMy (§ IV), we now look at how the spatial and temporal availability of tags can be further exploited to create alternative uses. Using Figure 4 as a guide: (a) In an area with multiple finders, using a *single* tag in a fixed location we can estimate how crowded a monitored area is (§ VI-A), while using *multiple* tags in fixed locations we can study properties of their flow (§ VI-B). (b) Targeting a single finder, using a *single* tag, placed for a short period in proximity with a target, and have this finder move to a commonly visited place, we (or an attacker) can estimate where this target finder could have gone from a list of possible destinations (§ VII-A). (c) Again, targeting a single finder, but this time using *multiple* tags, each placed along any arbitrary area (or paths), we (or an attacker) can estimate the trajectory taken by this finder (§ VII-B). (d) Using multiple tags (or simply emulating multiple tagIDs with a single transmitter) and any arbitrary number of finders, we can encode a message into the sequence these tagIDs are transmitted. As we will discuss, in (b), (c) and (d) the respective finder(s) are unaware of the respective use. Currently, only disabling Bluetooth or opting out of the FindMy service can mitigate this issue.

VI. Crowd Monitoring

We now evaluate how well smart tags on Apple's FindMy network can be used for *crowd monitoring*. We first present our results for crowd size estimates, evaluated against a state-of-the-art image recognition approach. Next, we present results for crowd flow which we evaluate against passive measurements of commonly used Wi-Fi management frames [2].

A. Crowd Size – Using a single tag

For this evaluation, we conducted 8 measurements, of 3 hours each, from July to September 2021 in a large public square in the city of Munich, Germany. During these months, this main square is often crowded due to shops, restaurants and metro stations nearby. Our smart tag setup consisted of an ESP32, advertising at ∼1 second intervals and using high TX power (+9 dBm) for maximum discoverability. We evaluate these measurements against the people count obtained by image recognition, which we describe next.

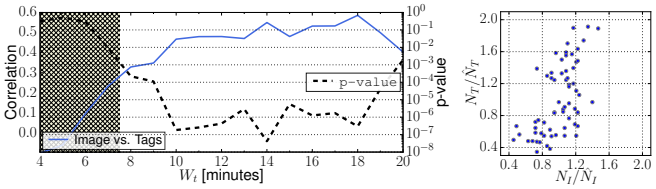
1) Image recognition: The use of images for crowd estimates produces some of the most accurate results with the use of inexpensive hardware [2]. Modern approaches based on Convolutional Neural Networks have quickly become the state of the art for all image recognition tasks, and in spite of their capabilities, such methods are not extensively used due to privacy concerns raised by their usage. Those concerns include regulatory legislation in several countries. For our evaluation, we used images from a publicly available web-cam[2], openly streaming images at 5 seconds per frame, with a 2048x1536 resolution. For that, we use the `Mask R-CNN` [18], from which we extract the total count of persons per frame. `Mask R-CNN` performs multi-class object instance segmentation, detecting and dividing each class instance in the prediction. This segmentation is of key importance for the detection of people in a large environment since they tend to stay in groups. `Mask R-CNN` is able to detect different people that overlap each other, increasing the accuracy of the model [18].

2) Size Measurements Description: For crowd size analysis, we correlate the total number of persons identified using image recognition against our smart tags approach. From the latter, we identify a unique device using the time a set of reports was uploaded to iCloud (see § IV).

3) Results: To best estimate the time window to aggregate tag reports, we correlated the number of identified finders over different time window (W_t, or bin sizes). Figure 5a depicts how the Pearson correlation value changed with bin

[2]https://www.ludwigbeck.de/webcam (will be **removed** in camera-ready)

sizes, while it also shows the `p-value` for those sizes. The `p-value` estimates the probability the estimated correlation coefficient was due to randomness, and we adopt `p-value` < 0.001 as our confidence interval, below which results are deemed acceptable. From the analysis, we note that only from W_t of 8 minutes we obtained `p-value` below the confidence interval, and the correlation coefficient reaches its maximum value at 18 minutes, with a value of 0.58 which corresponds to a strong correlation between both values. The highest values around 15 minutes could be explained by the expected time an iPhone takes to upload location reports (see § IV). Figure 5b depicts the best relationship between the normalized values for tags (N_T) and from images (N_I), at W_t 18 minutes. Thus, coarse-grained crowd size monitoring with a modest time lag appears feasible.



(a) Correlation and bin sizes (W_t) (b) $W_t = 18$ mins

Fig. 5: Relationship between numbers from tags (N_T) and from images (N_I), for different bin sizes (W_t).

B. Crowd Flow – Using multiple tags

We now evaluate how smart tags could be used to study the flow of a crowd. That is, we explore how well we can measure moving time and waiting time (or dwell time) using a pair of smart tags in a large urban environment. We compare our method with measurements done using Wi-Fi management frames, as those are currently widely adopted by researchers and commercial applications [2].

1) Wi-Fi Management Frames: Mainly due to its simplicity and reported accuracy [2], [19], crowd monitoring using Wi-Fi management frames is extensively used. In principle, all Wi-Fi enabled devices send a series of management frames, often used to search for available access points and to establish/maintain existing connections. These frames contain a device identifier (MAC), which can, in turn, be tracked through space and time while uniquely identifying a mobile device. To mitigate this traceability, since 2014, mobile devices perform MAC randomization at ever increasing rates and new schemes [20]. For our purposes, however, discarding locally managed addresses and subsampling our measurements with only global addresses suffices for our first order approximations of time between vantage points. From our measurements, on average, 26% of management frames captured were from non-random MAC addresses. We measured this using a Raspberry Pi 3, with two external antennas, hopping between the non-overlapping channels 1, 6 and 11.

2) Flow Measurements Description: For crowd flow analysis, we compare the distribution of time intervals a set of devices takes to be observed between two vantage points. These points were 176 meters apart and were chosen at the city center of Munich, Germany, in a commercial area where only pedestrians are allowed. We conducted 3 measurements of 2 hours each on 6/7/8 September 2021. From the measurements of the smart tags, we identify a unique device using the time a set of reports was uploaded to iCloud (t_r, see § IV). If such a bundle of reports contained at least one record at each vantage point, we could then infer the time interval the device took between both locations. Similarly from Wi-Fi frames, this interval corresponds to the time between consecutive records at each observed location.

3) Results: The left panel on Figure 6 shows the histogram of measured times between vantage points, with a mode at 2 minutes from both sources. Furthermore, to meaningfully analyze our measurements, we decompose them into log-normal distributions, using the widely used Gaussian-Mixture Model. This unsupervised learning method decomposes an input set into a pre-determined number of Gaussians. To select the best number of clusters, we used the BIC method [21] which estimates how well a given model explains the variance in the measured value. Our empirical results suggest that the ideal number of clusters for the Wi-Fi and smart tags measurements is 3. Following this classification, the right panel on Figure 6 depicts the similarities in the distributions of the estimated walking and waiting times from tags and Wi-Fi. The average walking time was 2.41±0.04 minutes using tags and 2.28±0.04 minutes using Wi-Fi. That yields an equivalent ∼4.5 km/h walking speed, in line with existing urban pedestrian research [22]. The average estimated waiting time (assuming the mean walking time above) was 19.33±1.44 minutes using tags and 20.54±0.69 using Wi-Fi. A possible interpretation of these values is the expected time pedestrians have spent at shops along the way.

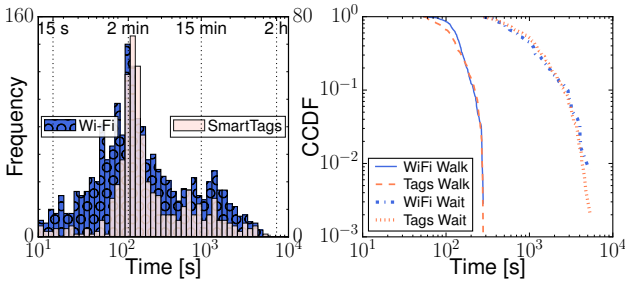


Fig. 6: Crod Flow [Left] Time between vantage points. [Right] Estimated walking and waiting times between vantage points.

Takeaway (§ VI): Bluetooth trackers can be used for crowd monitoring, with comparable results to widely used alternative solutions. These alternatives, however, may disclose personal information and always generate data that needs to be handled with care, such as personal identities. For crowd sizes, estimates with *a single tag* strongly correlate with estimates using state of the art image recognition. Furthermore, different aspects of crowd flow were accurately estimated using multiple

tags when compared to Wi-Fi measurements. In both use-cases, our approach always guarantees the anonymity of the studied subjects given the reporting mechanism and end-to-end encryption of this Apple service.

VII. Deliberate Tracking

In this section, we present proof of concept (PoC) evaluations and possible mitigation strategies to information being leaked by the Apple FindMy service. We demonstrate how this leakage may allow an attacker to track a victim's device through *timing attacks*, enabled by the reporting system implemented by Apple. The experiments we present used only our own devices to avoid disclosing unwanted information from other subjects. Apart from privacy concerns, it was necessary to have control of the phone's settings and times when connected to Wi-Fi or cellular network.

Overview: We demonstrate two examples of information leakage: A) Destination inference, in which the timing between sensing a tag and uploading a report may disclose where a victim could have gone; B) Path reconstruction, in which a sequence of visited places can be precisely inferred. Both examples rely on the bundling of reports (see § III) as well as the difference in uploading delay when connected to different networks (see § IV). *Note that the threats we present only require the victim being near a tag for a brief period of time, and not being tracked by inadvertently carrying a tag.*

A. Remote Destination Inference – Using a single tag

This attack relies on the timing of location reports, but precisely the difference between sensing a tag (t_c) and uploading a report (t_r). Furthermore, the modulation of the TX power can limit the range of BLE beacons, helping ensure only a victim's phone is affected.

1) Threat Model: An attacker, who wants to know where a victim has gone after an encounter, performs a timing attack using one tag. The attacker knows the victim's most visited locations, and the victim is only connected to cellular while outdoors and connects immediately to Wi-Fi when reaching her destination. During the encounter, the attacker "tags" a victim's phone by transmitting a series of beacons. The victim's phone, while on cellular, will store the reports until reaching her destination where, on Wi-Fi, it will upload all location reports for the attacker's tag. With the difference between sensing and uploading the reports, an attacker can limit (or pinpoint) the most likely destination of the victim. The victim is unaware this attack is tracking place, and only disabling Bluetooth or the FindMy service can mitigate it.

2) PoC Description: For this, we used an iPhone 12 (iOS 15) and one tag, configured at -6 dBm ensuring only at close proximity our phone would sense our tag. We enabled our tag next to our iPhone for 1 minute, then moved 18.5 km (11.5 miles) to a destination, where we finally enabled Wi-Fi.

3) Results: Our moving time was ∼29 minutes, and the difference between t_c and t_r was ∼35 minutes. Furthermore, we observed similar behavior on sensing and immediately uploading reports once on Wi-Fi, as previously discussed.

B. Path reconstruction – Using multiple tags

Given the bundling of reports (see § III), intentionally positioned *lost* tags can form a sequence of "breadcrumbs" which can then disclose the path followed by a phone. Similarly to the previous example, TX power can be modulated to ensure shorter coverage from each tag. As a proof of concept, we conduct one experiment.

1) Threat Model: An attacker, willing to find out the whereabouts of a victim through an area of interest, places tags at known locations. This attack relies on the victim having her phone connected to the cellular network only while moving and eventually connecting to Wi-Fi after the monitored journey. The victim's phone will then sense these tags, keeping the order of the observed tags. Once uploaded, the location reports disclose where and when the victim had been. The unique t_r, appended by iCloud when receiving a bundle of reports, uniquely identifies a finder device (see § III). The victim is unaware this attack is taking place, and only disabling Bluetooth or the FindMy service can mitigate it.

2) PoC Description: For this example, we used 3 iPhones (7 and 8 on iOS 14.8 and 12 on iOS 15), and placed 3 pairs of tags in a straight line, with each pair at 150 meters away from the next pair, configured at -6 dBm to ensure that only at close proximity devices would sense our tags. We stayed for 5 minutes, at a distance of 2 meters from each pair of tags, then walked to the next location (in ∼2 minutes). We disabled Wi-Fi until reaching a planned location away from the tags to ensure it did not unexpectedly upload any reports.

3) Results: From all three phones, we were able to reconstruct the path and timing taken at each location. Figure 7 depicts the transitions between each state (static or mobile) as well as the corresponding t_c contained in each location report. We also noted that, on all phones, the upload of reports happens within the first 5 minutes of switching to Wi-Fi from cellular-only. With such information, an attacker can reconstruct the set of visits of a victim and obtain an accurate estimation of the time spent at each place. However, if a finder device uploads a bundle of reports before all visits are done, then the attacker will not be able to fully reconstruct a trajectory.

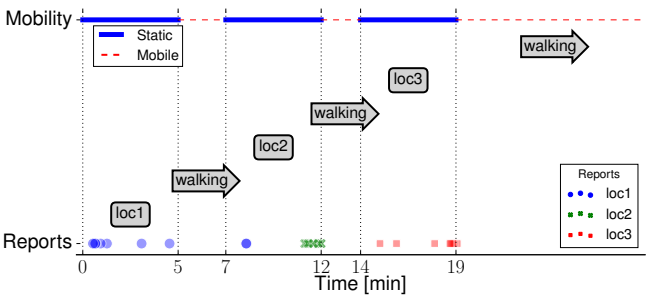


Fig. 7: Path reconstruction.

C. Mitigation

We now discuss mitigation strategies applicable to both path reconstruction and destination inference cases. Until Apple addresses these issues, users can only disable Bluetooth or the FindMy services to prevent their location information unknowingly being leaked, impairing the functionally of the service. From the system's perspective, providers like Apple could (1) reduce the granularity of the *received* timestamps (t_r) or remove them altogether from the location reports, (2) randomize when reports are uploaded, no longer determined by the connectivity available, or (3) initiate the uploading of reports after moving a random distance. None of these systems solutions should impact the functionality of the service, while still protecting the privacy of its users.

VIII. TagComm – Covert Channel Using BLE Trackers

Now, we turn to an illustration of a side-channel communication, built on the FindMy service, which we name *TagComm*. In this section, we delve into the design a simple unicast protocol, which essentially uses the sensing of tags by nearby iPhones to encode and transport information. Figure 8 outlines our proposed design: we create an artificial tag that changes its beaconed tag ID over time (chosen from a pre-defined alphabet, encoding messages as sequences of the transmitted tag IDs, without the awareness of any nearby finder. Our proposed system could, for example, be used by an attacker trying to exfiltrate data from an air gapped system (cf. [8]). Note that, the transmission is end-to-end encrypted as neither the finder nor Apple are able to decode that any specific tag ID is being transmitted. The decoding of the IDs transmitted, and therefore the final message, is only possible by the owner as discussed in Section III.

Overview: Our protocol uses a set of tag IDs and their permutations to encode information. That is, for N available tags we can encode $\lfloor \log_2(N!) \rfloor$ bits of information. Additionally, we include a set of header bits as well as a parity bit to be encoded along with the message payload. These extra bits and a pre-defined number of tags guarantee a transmitted message can *eventually* be recovered, as will be presented next.

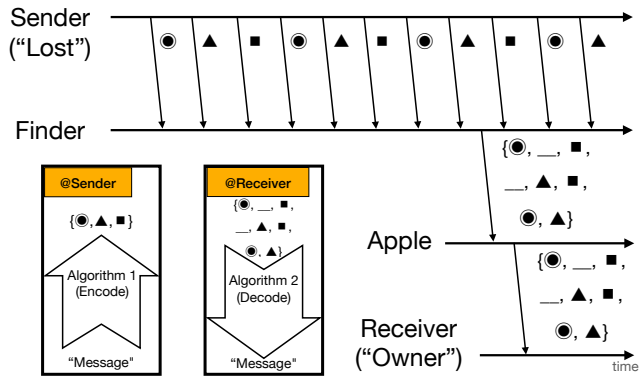


Fig. 8: *TagComm* protocol example, encoding a message as a sequence of tag IDs, silently and securely transmitted by a *finder*.

A. Encoding

In order to maximize the amount of information being sent and provide basic integrity guarantees, we use the permutations of N tag IDs to encode the information we want to transmit. Furthermore, the understanding of the sensing and reporting behavior from Section IV establishes bounds to how fast information can be transmitted.

Algorithm 1: Encoding input word into sequences

```
Input: S,w; /* Symbols and word encoded */
Output: E;  /* Encoded sequence */
L ← length(S) ;  /* Length of S */
assert(L! >= w) ;  /* From ❶ */
E ← [ceil(w/(L − 1)!)] ;  /* One item list */
for idx ∈ range(L − 1, 0, −1) do  /* From ❷ */
    w −= (E[−1] − 1) ∗ idx! ;  /*E[−1]: last */
    e = ceil(w/(idx − 1)!);
    E.append(e);
```

1) Input to sequence of symbols: Given an input message to be transmitted W and a set of encoding symbols of size N, we iteratively divide the interval of N! to find the corresponding sequence to be used. Note that this requires W < N! (as ❶). For example, for N=5 and W=42, we define an order for the resulting set of symbols, *i.e.*, a<b<c<d<e (❷). Next, we split the interval 5! into 5 equally sized blocks, as depicted in Figure 10a. As 42 is found within the second block, the symbol *b* is removed and set as the first symbol. These steps are repeated until all symbols have been removed, yielding the final sequence *bdeca*. This procedure allows us to encode $\log_2(N!)$ bits of information using N tags. Algorithm 1 systematically describes these steps.

2) Defining N=16: Given ❶, we define the code efficiency as the ratio $\lfloor \log_2 N! \rfloor / \lceil N \log_2 N \rceil$ (as ❷). That is, the maximum number of bits encoded by the minimum number of bits required for all used symbols N. Given these observations, Figure 9 shows the variation in ❷ for different values of N up to 20 tags[3], with its highest efficiency at N=16, which we use in our experiments. This leaves us with a total of 44 bits, and their use will be further described next.

3) Frame and supporting bits: To ensure the integrity of the information being transmitted and to allow the receiver to decode that information, we define a simple frame to our protocol, illustrate in Figure 10b. Three *header* bits (as MSB) distinguish different message types: STX (0b000) the start of transmission, F0 and F1 alternating even and odd frames (0b010 and 0b100, respectively), and EOT end of transmission (0b110). These bits ensure the receiver can deterministically identify the start of a transmission, new frames as well as the end of a transmission. This guarantee

[3]Largest number of permutations that fits in 64 bits.

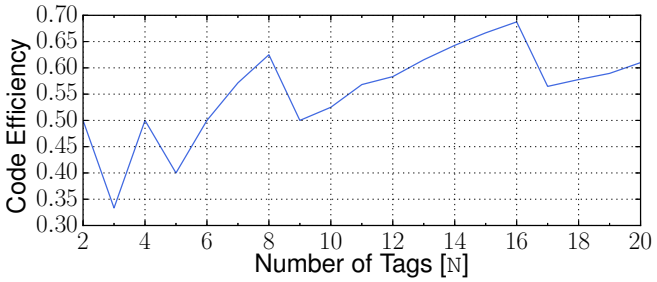


Fig. 9: Code efficiency given the number of symbols (tag IDs) being used to endcode a message, with its maximum at 16.

is achieved by ensuring the initial symbol of a sequence will be unique for each frame type, as a consequence of the values chosen for the header bits. Additionally, a parity bit (as LSB) provides a minimal "checksum" to the message being transmitted once its decoded. Finally, the remaining bits are used for the message payload, which in our setup, consists of 40 bits.

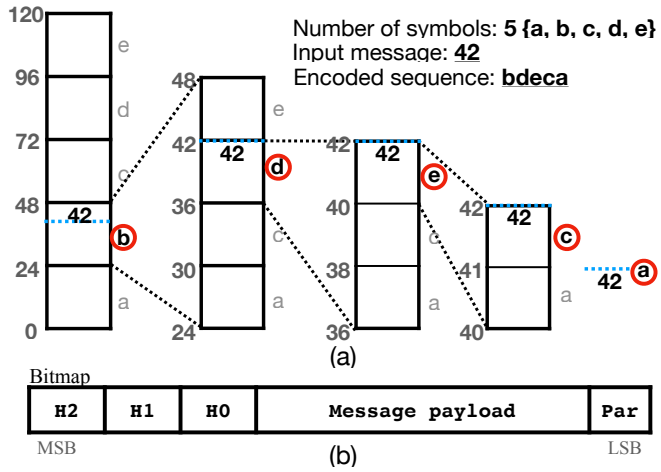


Fig. 10: Protocol definitions. (a) Encoding a value as a sequence of symbols. (b) Frame bitmap.

4) Transmission integrity: To transmit a block of information `W` using Tagcomm, a node will chunk it into words `w` of 40 bits. A transmission will start with an `STX` frame (*i.e.*, header bits set to `0b000`) that will carry the total number of words to be expected in its payload. Next, each word is encoded as a sequence (as described above) as alternating frames of type `F0` and `F1`. Finally, a transmission is terminated with an `EOT` frame. Note that, as the protocol requires all symbols to be transmitted, if a receiver is unable to reconstruct an entire sequence, the corresponding word `w` cannot be retrieved. For details on error recovery, see Section VIII-B.

B. Decoding

Once received by the owner device, location reports for a series of tag IDs should be decoded into its original message. Essentially, this is done by inverting the steps done while encoding a frame into a sequence. Once the different frames are decoded and parity bits verified, the original message can finally be reconstructed.

1) Tags alignment: As the duration of each tag being transmitted is predefined (*e.g.*, t_{tag}), the first step in decoding a message is aligning each received tag in *slots* of size t_{tag}.

2) Frames alignment: As discussed in Section VIII-A, each frame type starts distinct symbols, represented and transmitted in TagComm as tags. Similar to the tags alignment, the duration of how long a frame is sent is also predefined, for example 5 minutes. This way, knowing the expected sequence of frames, *i.e.*, `STX`, `F0`, `F1`, ..., `EOT` (as ❸), and their corresponding starting symbols, we can align all received frames and start the decoding step.

3) Decoding frames: Once the sequences that encode each frame are identified, we can decode the information by reversing the steps explained in Section VIII-A. That is, assuming a predefined order between tags (*e.g.*, ❷), and taking the encoded sequence as input, we can recover the initial message by adding up the partial contributions each symbol had in splitting the `N!` interval, as described in Algorithm 2. After decoding, we then verify the presence of errors in the next step.

Algorithm 2: Decoding sequences into words

Input: S, E; /* Symbols and encoded seq. */
Output: w; /* Decoded word */
$L \leftarrow$ length(S) ; /* Length of S */
$P \leftarrow [\,]$; /* Empty list */
for $i, e \in$ *enumerate*(E) **do**
  idx $\leftarrow S$.index(c) ; /* Symbol e index */
  P.append$((L-1-i) *$ idx); /*Partial sum*/
  S.pop(idx)
$w \leftarrow$ sum(P) ; /* Add up all partials */

4) Error Correction: Once each frame has been decoded from the input sequences, we can validate the integrity of the frame with its parity bit. Furthermore, the expected sequence of frame types (*i.e.*, ❸), combined with a parity bit, allows us to recover messages when a single tag (out of a sequence) is not received. The position of the missing tag can be determined when aligning the tags in each frame (see above), and finally verified with the corresponding parity bit, allowing us to reconstruct the orginal message in the next step.

5) Final message reconstruction: Once all frames have been decoded, the total number of expected frames sent as the payload of `STX` frames can be read and verified. Finally, all the information encoded in a series of sequences of tags can be reconstructed.

C. TagComm Experiment

In this section, we describe the set of experiments we conducted to test our TagComm system. Furthermore, we present a series of observations made from the results obtained.

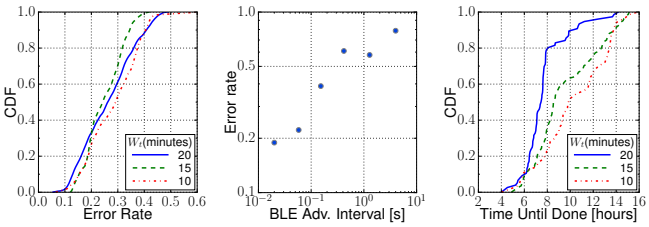
D. Setup

For all measurements, we transmitted a set of 10 randomly generated words of 40 bits. These were transmitted using the protocol described in Section VIII. As previously discussed, we used a single ESP32 as our lost tag, which implemented the transmitter/encoder logic (see § VIII-A. As a *finder*, we had an iPhone 12 (iOS 15) nearby, connected to Wi-Fi at all times. As discussed in Section III, this setting allows us to estimate a best-case scenario given the expected frequency the iPhone would publish location reports for our tag (*i.e.*, within 15 minutes more than 50% of the time).

E. Results

We were able to successfully transmit a set of random words, as described above. To better verify the limits of our system, we varied some of the parameters, such as word duration and BLE advertisement interval.

Definitions: For this analysis, we define the error rate as the fraction of tag slots during which no reports were received. That is, given the tag slot duration (*e.g.*, 30 s), the error rate of a transmission corresponds to tag slot intervals during which a finder was present but no report was sent. Further, we define the time until done (TUD) as the expected minimum time required to decode a complete message, from starting the transmission until it is fully decoded by the owner's device.



(a) (b) (c)

Fig. 11: Error rate and TUD for different settings. (a) CDF of error rate and different frame duration (W_t). (b) Error rate and BLE advertisement intervals. (c) CDF of TUD and W_t.

Frame duration and error rate: The different frame durations we tested produced small differences between error rates. We observed 26.17% for 20 minutes, 24.73% for 15 minutes, and 28.82% for 10 minutes on average, and their distributions are depicted in Figure 11a. This indicates that repeating a tag at a certain position for longer periods of time does not affect the probability it will be detected by a Finder.

Larger BLE advertisement intervals increase error rate: We compare the effect different Bluetooth LE advertising intervals had in the error rate. From our measurements, we note that for larger advertising intervals, less tags were observed per unit of time. Figure 11b depicts the monotonic increase in error rate with increased adv. intervals. This could be explained by the probabilistic nature of these intervals, and to which transmitters do not have any control [16].

Frame duration and time until done: We tested frame window sizes around 15 minutes (10, 15, 20), as previous experiments showed that to be the expected time over 50% of reports take to be published. We measured each configuration for 72 hours, and computed expected values for TUD from 100 different random starting points in each setting. For these measurements, we observed 7.73±0.22 hours with 20 minutes, 9.57±0.30 hours with 15 minutes, and 10.36±0.36 hours with 10 minutes. The distributions of TUD for each configuration is depicted on Figure 11c. Interestingly, using two iPhones on the same iCloud account and placed near a tag, did not produce statistically significant improvements.

F. Mitigation

Currently, users can avoid inadvertently transporting information with a similar system by disabling Bluetooth on their phones or opting out of the FindMy services. From the system side, while keeping the main functionality of the FindMy services, Apple can limit the number of updates issues by a finder, as well as limit the number of available reports per tag or decrease the accuracy of the time stamps used. Notably, Apple currently does not notify users about one of these crafted tags being around, as we did not get a single notification during our experiments, using multiple iCloud accounts.

IX. Discussion

Privacy: To preserve privacy of the individuals part of our crowd sensing experiments, we (1) discard all original MAC addresses from the Wi-Fi measurements, leaving records that can no longer identify the owners of the original devices, (2) we compute the metrics from each image and store only the counts per frame, using publicly available images, and (3) used our own equipment to demonstrate the possible information leakage from the FindMy services. For the communication experiments, by using our own devices and iCloud accounts, we ensure the privacy and resources of other individuals were not affected by our work. However, our work unveils possible attack vectors which could be exploited, compromising the security and privacy of the subjects involved.

Ethics: Our measurements and analysis were designed and executed to minimize exposing information about subjects being studied. Whenever possible, we limited our study to our own devices, and when studying crowds we discarded all identifiable information. However, we understand the methods presented could be used in other unintended ways. Therefore, we believe such study may contribute to the design of future versions of Bluetooth tracking systems.

Crowd Sensing: Our analyses show acceptable results using *a single tag* on the Apple FindMy service to sense aspects of a crowd. More importantly, our system provides privacy guarantees when used with a large group of subjects. Unlike in the deliberate tracking examples, a large group of unknown individuals ensures no single subject can be identified or have further information disclosed.

Extended support to other platforms: To enable further studies with such smart tags system, we extend the support originally implemented by Heinrich *et al.* [1] to `macOS` (*e.g.*, developing of new applications and debugging) and the Amazon Echo [3], [4] (*e.g.*, home IoT turn into sensing device) to be used as a tag.

More Finders may increase Tagcomm delivery guarantees: During our Tagcomm experiments, using a single extra Finder did not yield significant improvements in the reliability of sending messages. For applications deployed in spaces where multiple finders could be passing by may increase further the guarantees messages are transmitted.

X. Conclusion

In this paper, we present how Bluetooth trackers (or tags) can be used beyond their originally designed purpose, of tracking lost devices. We show how crafted tags can be used as crowd sensing devices, with relative estimates of large groups of people. These estimates are size with even a *single* tag and flow using multiple tags along a monitored path. Furthermore, we demonstrate through a series of controlled experiments how the Apple FindMy service currently discloses sensitive location information from passive finder devices. An attacker may, in turn, reconstruct a victim's path and visits as well as a possible final destination, currently exposing billions of Apple devices [1].

Furthermore, we present Tagcomm, a proof-of-concept out-of-band communication channel using Apple tags. Using a simple protocol, we demonstrate how various tag IDs can be used to encode any arbitrary information and transmitted over a secure end-to-end encrypted channel, without the knowledge of the phones that handle part of this communication path. Our intent is to raise awareness of such possibility, while discussing possible uses which include side-channel communication that could leak sensitive information from a compromised system.

Future iterations of our work will consider other Bluetooth trackers for crowd sensing, and leverage TagComm to estimate users' behavior. Furthermore, similar systems providing raw reports (*i.e.*, not aggregates over time) will be verified for the vulnerabilities presented here.

Reproducibility: To foster further research with TagComm, we make our code and sample data openly available [4] along with an extended support for other platforms to be used for either communication or sensing [3].

References

[1] A. Heinrich, M. Stute, T. Kornhuber, and M. Hollick, "Who can find my devices? security and privacy of apple's crowd-sourced bluetooth location tracking system," *Proc. Priv. Enhancing Technol.*, vol. 2021, no. 3, pp. 227–245, 2021. [Online]. Available: https://doi.org/10.2478/popets-2021-0045

[2] A. Draghici and M. van Steen, "A survey of techniques for automatically sensing the behavior of a crowd," *ACM Comput. Surv.*, vol. 51, no. 1, pp. 21:1–21:40, 2018. [Online]. Available: https://doi.org/10.1145/3129343

[3] L. Tonetto *et al.*, "Tags Crowd Sensing," https://home.in.tum.de/~tonetto/smarttag_artifacts.tar.bz2, 2022, [Online; accessed Jan-2022].

[4] L. Tonetto, "Tagcomm," https://home.in.tum.de/~tonetto/tagcomm_artifacts.tar.bz2, 2022, [Online; accessed Jan-2022].

[5] J. Martin *et al.*, "Handoff all your privacy–a review of apple's bluetooth low energy continuity protocol," *Proceedings on Privacy Enhancing Technologies*, vol. 4, pp. 34–53, 2019.

[6] J. Martin, D. Alpuche, K. Bodeman, L. Brown, E. Fenske, L. Foppe, T. Mayberry, E. C. Rye, B. Sipes, and S. Teplov, "Handoff all your privacy - A review of apple's bluetooth low energy continuity protocol," *Proc. Priv. Enhancing Technol.*, vol. 2019, no. 4, pp. 34–53, 2019. [Online]. Available: https://doi.org/10.2478/popets-2019-0057

[7] M. Weller, J. Classen, F. Ullrich, D. Waßmann, and E. Tews, "Lost and found: stopping bluetooth finders from leaking private information," in *WiSec '20: 13th ACM Conference on Security and Privacy in Wireless and Mobile Networks*. ACM, 2020, pp. 184–194. [Online]. Available: https://doi.org/10.1145/3395351.3399422

[8] A. Dorais-Joncas *et al.*, "Jumping the air gap: 15 years of nationstate effort," https://www.welivesecurity.com/2021/12/01/jumping-air-gap-15-years-nation-state-effort/, 2022.

[9] M. Guri, B. Zadov, and Y. Elovici, "Led-it-go: Leaking (A lot of) data from air-gapped computers via the (small) hard drive LED," in *Detection of Intrusions and Malware, and Vulnerability Assessment - 14th International Conference, DIMVA 2017, Bonn, Germany, July 6-7, 2017, Proceedings*, ser. Lecture Notes in Computer Science, M. Polychronakis and M. Meier, Eds., vol. 10327. Springer, 2017, pp. 161–184. [Online]. Available: https://doi.org/10.1007/978-3-319-60876-1_8

[10] M. Eichelberger, S. Tanner, G. Voirol, and R. Wattenhofer, "Receiving data hidden in music," in *Proceedings of the 20th International Workshop on Mobile Computing Systems and Applications, HotMobile 2019, Santa Cruz, CA, USA, February 27-28, 2019*, A. Wolman and L. Zhong, Eds. ACM, 2019, pp. 33–38. [Online]. Available: https://doi.org/10.1145/3301293.3302360

[11] M. Guri, Y. A. Solewicz, A. Daidakulov, and Y. Elovici, "Acoustic data exfiltration from speakerless air-gapped computers via covert hard-drive noise ('diskfiltration')," in *Computer Security - ESORICS 2017 - 22nd European Symposium on Research in Computer Security, Oslo, Norway, September 11-15, 2017, Proceedings, Part II*, ser. Lecture Notes in Computer Science, S. N. Foley, D. Gollmann, and E. Snekkenes, Eds., vol. 10493. Springer, 2017, pp. 98–115. [Online]. Available: https://doi.org/10.1007/978-3-319-66399-9_6

[12] A. M. Al-Shaery, S. S. Alshehri, N. S. Farooqi, and M. O. Khozium, "In-depth survey to detect, monitor and manage crowd," *IEEE Access*, vol. 8, pp. 209 008–209 019, 2020. [Online]. Available: https://doi.org/10.1109/ACCESS.2020.3038334

[13] C. A. Pouw, F. Toschi, F. van Schadewijk, and A. Corbetta, "Monitoring physical distancing for crowd management: Real-time trajectory and group analysis," *PloS one*, vol. 15, no. 10, p. e0240963, 2020.

[14] S. Chang, E. Pierson, P. W. Koh, J. Gerardin, B. Redbird, D. Grusky, and J. Leskovec, "Mobility network models of covid-19 explain inequities and inform reopening," *Nature*, vol. 589, no. 7840, pp. 82–87, 2021.

[15] Apple, "FindMy App," https://www.apple.com/icloud/find-my/, 2021, [Online; accessed September-2021].

[16] Á. Hernández-Solana, D. P. D. Cerio, A. Valdovinos, and J. L. Valenzuela, "Proposal and evaluation of BLE discovery process based on new features of bluetooth 5.0," *Sensors*, vol. 17, no. 9, p. 1988, 2017. [Online]. Available: https://doi.org/10.3390/s17091988

[17] Apple, "Accessory Design Guidelines - 37.5 Advertising Interval," https://developer.apple.com/accessories/Accessory-Design-Guidelines.pdf, 2021, [Online; accessed September-2021].

[18] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 2980–2988. [Online]. Available: https://doi.org/10.1109/ICCV.2017.322

[19] L. Tonetto, M. Untersperger, and J. Ott, "Towards exploiting wi-fi signals from low density infrastructure for crowd estimation," in *Proceedings of the 14th Workshop on Challenged Networks (CHANTS)*. ACM, 2019, pp. 27–32. [Online]. Available: https://doi.org/10.1145/3349625.3355439

[20] E. Fenske, D. Brown, J. Martin, T. Mayberry, P. Ryan, and E. C. Rye, "Three years later: A study of MAC address randomization in mobile devices and when it succeeds," *Proc. Priv. Enhancing Technol.*, vol. 2021, no. 3, pp. 164–181, 2021. [Online]. Available: https://doi.org/10.2478/popets-2021-0042

[21] G. Schwarz, "Estimating the dimension of a model," *The annals of statistics*, pp. 461–464, 1978.

[22] R. L. Knoblauch, M. T. Pietrucha, and M. Nitzburg, "Field studies of pedestrian walking speed and start-up time," *Transportation research record*, vol. 1538, no. 1, pp. 27–38, 1996.

Publication 7

©2022 Springer Nature, reprinted with permission from:

Kister, P., Tonetto, L. On the importance of structural equivalence in temporal networks for epidemic forecasting. Sci Rep 13, 866 (2023). https://doi.org/10.1038/s41598-023-28126-w

RightsLink

On the importance of structural equivalence in temporal networks for epidemic forecasting

Author: Pauline Kister et al
Publication: Scientific Reports
Publisher: Springer Nature
Date: Jan 17, 2023

Copyright © 2023, The Author(s)

Creative Commons

This is an open access article distributed under the terms of the Creative Commons CC BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.

To request permission for a type of use not listed, please contact Springer Nature

© 2024 Copyright - All Rights Reserved | Copyright Clearance Center, Inc. | Privacy statement | Data Security and Privacy | For California Residents | Terms and ConditionsComments? We would like to hear from you. E-mail us at customercare@copyright.com

Title: On the importance of structural equivalence in temporal networks for epidemic forecasting
Authors: Pauline Kister (TUM), **Leonardo Tonetto** (TUM)
Journal: Nature Scientific Reports (13-1)
Publishing date: January 17, 2023
Reference: [7]

Publication Summary

In this paper we discuss the use of machine learning models and node embeddings to analyze the spreading process of infectious diseases in a population. We evaluate the impact of homophily (similarity between nodes based on shared characteristics) and structural equivalence (similarity between nodes based on their position in the network) on the prediction of disease spread. We find that embeddings balanced towards preserving structural equivalence outperform those focused on homophily, indicating that structurally equivalent nodes behave similarly during an epidemic. This finding has implications for predicting future epidemics and assessing the risk of infection for different groups in the population.

We highlight the importance of understanding disease spreading in complex networks like human society and the need for accurate predictions to develop containment strategies and vaccination plans. The paper discusses the concept of node embedding, which represents nodes in a network as low-dimensional feature vectors while preserving the network's structure. We use the `node2vec` algorithm, which conducts random walks through the graph, to compare embeddings with inward and outward exploration. We find that embeddings with outward exploration, prioritizing structural similarity, result in better predictions of epidemic dynamics than embeddings with inward exploration.

The results show that the predictive power of structural equivalence is stronger than the exact neighbors of a node, suggesting that a node's role in the network may be more relevant than its immediate connections for disease outbreak outcomes. We emphasize the significance of structural networks and weak ties in human contact networks, drawing on observations from sociology. We also discuss other embedding methods that focus on structural equivalence and cautionary results regarding the preservation of structural equivalence in `node2vec` embeddings.

In conclusion, the study demonstrates that embeddings preserving structural equivalence are more effective in predicting disease spread in human contact networks. This finding has implications for improving predictions of future epidemics and contact tracing efforts. By considering the behavior of structurally equivalent nodes, even when they are not directly connected, it becomes possible to assess the infection risk for different groups in the population.

Contribution

I came up with the idea and designed the study. Pauline Kister analyzed the data and wrote the base text describing the main observations. I wrote the manuscript. All authors reviewed the text.

scientific reports

Check for updates

OPEN

On the importance of structural equivalence in temporal networks for epidemic forecasting

Pauline Kister & Leonardo Tonetto✉

Understanding how a disease spreads in a population is a first step to preparing for future epidemics, and machine learning models are a useful tool to analyze the spreading process of infectious diseases. For effective predictions of these spreading processes, node embeddings are used to encode networks based on the similarity between nodes into feature vectors, i.e., higher dimensional representations of human contacts. In this work, we evaluated the impact of *homophily* and *structural equivalence* on `node2vec` embedding for disease spread prediction by testing them on real world temporal human contact networks. Our results show that structural equivalence is a useful indicator for the infection status of a person. Embeddings that are balanced towards the preservation of structural equivalence performed better than those that focus on the preservation of homophily, with an average improvement of 0.1042 in the f1-score (95% CI 0.051 to 0.157). This indicates that structurally equivalent nodes behave similarly during an epidemic (e.g., expected time of a disease onset). This observation could greatly improve predictions of future epidemics where only partial information about contacts is known, thereby helping determine the risk of infection for different groups in the population.

The recent outbreak of COVID-19 showed the importance of knowing the process of how a virus spreads in a population[1]. If the exact path a virus takes from person to person is known, better containment strategies, vaccination plans and preparations for future epidemics could be developed[6,13,25,28]. However, a process that might seem simple – one person infects the next – can quickly become complex in a network such as human society. The characteristic topology of a human contact network, including the scale-free distribution of node degrees and the formation of clusters, has an impact on the path a disease takes through these networks[4,14,16]. Furthermore, recent advances in human mobility and network epidemiology research allow us to expand our understanding of such complex problems. That is, if the temporal contacts network and attributes of the disease being studied are known, such as infection and recovery rate, it is possible to simulate the trajectory of epidemic outbreaks. Furthermore, machine learning models can facilitate the prediction of the remaining nodes, even if the disease attributes are not known and only partial information about a few infected nodes exists. One limitation of this approach is that most machine learning methods expect a set of feature vectors as input instead of a graph. That is, models are trained and validated using expected outcomes and not a representation of how nodes are connected over time. As a currently widely used solution, these learning methods rely on node embedding to convert the contact network into a set of feature vectors. The task of the node embedding is to represent each node in the network as a low-dimensional feature vector while preserving the structure of the graph as much as possible. This means that if two nodes are structurally similar, their resulting feature vectors are close to each other in the embedding space. In this way, information about node similarity can be preserved in this new representation. How good an embedding is depends on the context and the task it is used for. In the context of disease spreading, the embedding of a contact network should preserve all the information necessary for predicting *who* is infected, and *when*.

Common embedding methods used for disease spreading are based on the fact that a disease spreads from person to person, between nodes who are connected for sharing some kind of similarity, which corresponds to *homophily* in a network. Another relevant concept in network modelling is *structural equivalence*, where nodes with the same position in the structure of the network are infected at the same time. In real world networks, both of these notions of similarity appear at the same time and their influence for networks anaysis varies for different kinds of networks[10].

A popular embedding method is `node2vec`[10], an algorithm based on random walks. To discover and encode a graph's topology, `node2vec` conducts random walks through the graph. These random walks can be either

Technical University of Munich, Munich, Germany. ✉email: tonetto@in.tum.de

biased towards *inward* exploration while preserving homophily in the network, or towards *outward* exploration preserving structural equivalence. In this work, we used `node2vec` with inward, neutral and outward oriented random walks and compared predictions about the state of each node made by a logistic regression algorithm. This analysis reveals that the best predictions can be made with outward oriented random walks, which suggests that structural equivalent nodes show similar behaviour during the infection process, that is, better predictive models are obtained when structural equivalence is primed. In turn, these observations suggest that two or more structurally equivalent nodes in a network may be at similar risks of infection even if they are not directly connected, potentially helping contact tracing efforts during an infectious disease outbreak.

The influence of homophily and structural equivalence in a community has been observed by sociologists in many different contexts. In sociology, homophily in a network corresponds to *cohesion* or *strong ties* in society, while structural equivalence corresponds to *weak ties*. In some contexts, weak ties are more important than strong ties[9], which speaks for the importance of structural networks in human contact networks. In more recent studies, considering weak ties was helpful in different sociological contexts, such as the influence of indirect contacts on decision making[29], or the dismantling of organized crime[15]. In a more graph theoretical approach, researchers explored different methods of clustering people by their roles, relying on the fact that structurally equivalent nodes fulfill the same role in society[19].

While most node embedding methods focus on homophily, some have been developed to preserve only structural equivalence. The method proposed by Wang *et al.*[27] has a precise mathematical approach, while `struc2vec`[5] is based on random walks similarly to `node2vec`. Another popular embedding method uses the recursive nature of structural equivalence to create an embedding, as two nodes are considered structurally equivalent if their neighbors are structurally equivalent[26].

While reviewing node embedding methods, Junchen *et al.* confirmed that `node2vec` embeddings perform best against other methods at preserving local structure[12], whereas Schliski *et al.* included notes of caution by testing `node2vec` with different hyperparameter settings and concluded that `node2vec` does not preserve structural equivalence well, even with outward orientated hyperparameters[22]. However, these results should not affect our analysis as the authors focused at global structural equivalence, and we chose `node2vec` as it tries to achieve a balance between local structural equivalence and homophily.

Results

We observe that embeddings with outward exploration result in better predictions of epidemic dynamics than embeddings with inward exploration. That is, structural similarity has a stronger predictive power than the exact nearby neighbors of a node, or a node's role in a network may be more relevant than who their peers are for a disease outbreak outcome.

Our features used for prediction may consider either nodes with common neighbors to be similar, obtained through *inward* exploration, or nodes surrounded by an equivalent structure of neighbors to be similar, obtained through *outward* exploration (see "Methods" section). Furthermore, we evaluate our approach using the f1-score, which balances the accuracy of true and false predictions of positive and negative cases in all time steps of the outbreak considered (see "Evaluation Metrics" section). We run 50 simulations for each combination of dataset and SIR parameters, and report the averages and confidence intervals of the results.

We performed simulations on a total of 24 networks, with 6 of them derived from publicly available datasets, which are described in "Methods" section. By prioritizing *outward* explorations (*i.e.*, with a higher return parameter `p`, or a lower in-out parameter `q`), we obtain a better f1-score, as depicted in Fig. 1a and b. From our simulations, the best prediction results were achieved at $p = 50$ and $q = 0.01$. In comparison to the neutral values, outward exploration only achieved an average improvement in the micro f1-score of 0.01393 (95% CI −0.033 to 0.061), and inward exploration reduced the f1-score by 0.090211 (95% CI 0.037 to 0.143). Detailed results are presented in Tables 1, S5 and S6. In the best settings of parameters (*i.e.*, high `p` and low `q`), the resulting network structure representation focuses more on structurally equivalent nodes than on their immediate neighbors, meaning that prediction quality improves when structurally equivalent nodes are taken into account.

Additionally, some labels appear less frequently in our ground truth data than others and are therefore harder to predict. This resulted in a lower macro f1-score than the micro f1-score (see Fig. S2). Nevertheless, the prediction accuracy score increases similarly to the micro f1-score by 0.0185 (95% CI −0.029 to 0.066) for outward and

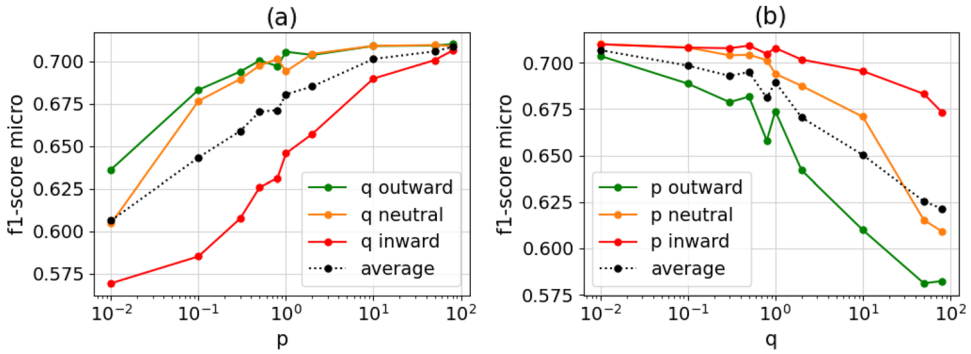


Figure 1. Comparison of inward, outward and neutral parameters.

	Micro f1-score (%)	Macro f1-score (%)
Neutral	69.4	52.9
Mean outward	70.8	54.8
Mean inward	60.4	40.2
Max outward	71.0	55.1
Max inward	67.5	50.0

Table 1. Mean and maximum scores for inward/outward parameters.

decreases by 0.1276 (95% CI 0.091 to 0.164) for inward exploration (see Table 1). With disease-specific metrics, such as outbreak size and prevalence, our results show that outward exploration yields better prediction accuracy than inward exploration. This observation is especially relevant regarding the prediction of final outbreak size and peak prevalence time, where it shows the most significant improvements (see Fig. 2). The prediction of the mean prevalence has the least improvement, with a difference of 0.0284 percentage points (95% CI 0.027 to 0.029). The effect of outward exploration can also be seen in the incidence, with an improvement of 0.0134 percentage points (95% CI 0.011 to 0.015) (see Fig. S5). The improvement of prediction accuracy seems to be the strongest early in the epidemic, showing a higher macro f1-score for the first 10% of time steps (see Fig. S9). This effect is however not reflected in the micro f1-score, where the difference between inward and outward exploration remains the same in different phases of the epidemic (see Fig. S8).

Interestingly, we note a small peak at the neutral values $p = 1$ and $q = 1$. If either $\mathtt{p}$ or $\mathtt{q}$ is set to 1, the prediction is either slightly better or worse than what could be expected from the values next to it, suggesting that the two parameters influence each other's outcomes. This behavior can also be seen in a 3D-plot of the scores for changing $\mathtt{p}$ and $\mathtt{q}$, as shown in Fig. S4. This unexpected behavior can only be seen in natural networks, where data were drawn from real subjects, as shown in Fig. S3. We conjecture that this behavior is likely caused by a characteristic of human contact networks that is not well modeled by artificially generated networks[2]. However, even for the real world networks this peak is small, and with outward exploration a more significant improvement can be achieved.

On the modelling and prediction of an infectious disease spreading, two nodes being *similar* means that their label (*i.e.,* predicted outcome at a given time) is more likely to be the same. That is, they will be infected and recover at the same time. Additionally, the state of a node in disease spreading will depend on its neighbors: a node is likely to be infected at the same time as its neighbors as they infect each other. Therefore, knowledge about the state of any neighbouring node can help the prediction accuracy, supporting the importance of considering homophily when studying disease spreading.

In a temporal network, the representation of nodes over time steps may not be ideal for predicting how an infectious disease will spread. As it is typically done, node i at time t and node i at $t + 1$ are connected by design of the supra-adjacency network. This representation is closer to reality since these nodes represent the same person at different time steps, *i.e.*, it is likely that they have the same state. However, unlike the task of predicting interests, when predicting the dynamics of an epidemic outbreak, a connection between two nodes is not necessarily caused by their similarity, but rather by both nodes being similar *because* they were in contact. In the corresponding temporal network, this observation results in a time delay, where the two nodes are similar during a single time step after they were in contact.

As shown by our results, structural equivalence can be indicative of the infection time of a person. People who are in the same structural position are likely to be infected around the same time. Additionally, communities might play a significant role in the spread of a disease, where if one community is infected and it has two neighboring communities, then all nodes that act as bridges between these communities are more likely to be

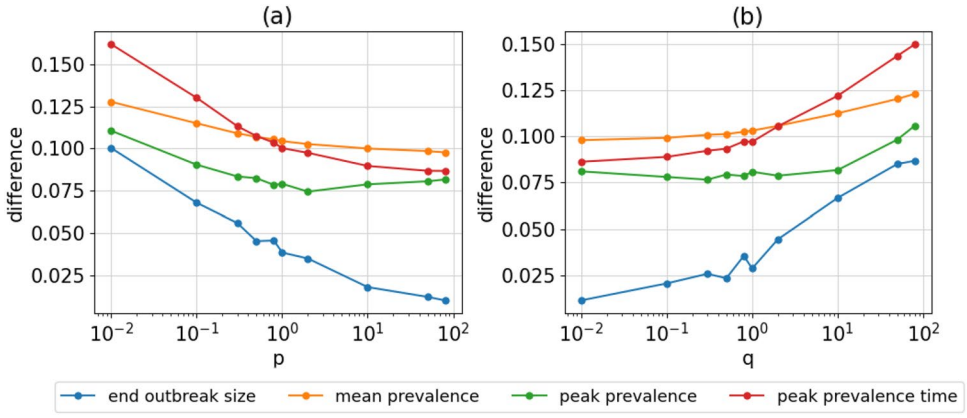


Figure 2. Difference in epidemiological metrics between prediction and simulation.

infected first, followed by the center nodes in the neighboring clusters, and then the nodes at the edges of these clusters. As a practical result, if we know the state of a node a neighboring cluster, we can then forecast the state for structurally equivalent nodes in the other clusters.

For the process of disease spreading, the balance of structural equivalence and homophily, *i.e.*, of outward and inward exploration in the random walk, allows us to draw conclusions about modelling the disease spreading process. Oftentimes, disease predictions follow the homophily approach – the disease is tracked from one person to the next. However, our results show that the prediction seem to work best with outward oriented parameters, and the prediction of a node's state is improved by taking into account the information about structurally equivalent nodes in the training set.

Discussion

Node embeddings are of paramount importance for the study and modelling of any large graph, including the forecast of how an infectious disease may spread. For this purpose, `node2vec` is able to preserve structural equivalence. This systematic approach allows us to compare the roles homophily and structural equivalence play in large temporal networks of human contacts. Since most infectious diseases, such as COVID-19[8,24], require close contact to spread from one person to the next, neighbors in a contact network will help drive how the disease spreads. In other words, the outcome of an outbreak depends on *when* and *how often* nodes are in contact[17,18]. Additionally, our results support the observation that the position of a person in their contact network seems to be indicative of infection time too.

Based on the results presented in this work, we note that it is possible to improve the prediction of a person's infection status at a given time by including information about structurally equivalent nodes. This information is encoded in our learning model through the inward and outward exploration parameters of the node embedding algorithm (*i.e.,* `node2vec`). Better prediction accuracy can help design countermeasures and prevent (or at least slow down) disease spreading. Structural equivalence in a human contact network is closely related to the roles people have in society. On a higher level, this might help identify groups of people that are infected first when only partial information about temporal contacts is available. Particularly, if a high amount of infections is detected within a certain group of people (*e.g.,* staff at hospitals or children at school), it is likely that a structurally equivalent group of people in other similar communities will be infected too.

Furthermore, our results highlight the importance of the availability of temporal contact data to study the spread of infectious diseases. Therefore, in preparation for future pandemics, contact tracing efforts could be established, while still protecting people's privacy as the highest priority. As human mobility tends to be highly predictable[23], it may be safe to assume even a sparsely sampled set of contacts prior to an outbreak could be representative. Therefore, yielding strong indicators of an uptick in infections by analyzing structurally similar nodes. Finally, to support and validate this conjecture and our observations, similar analyses are still required using temporal contact data from actual outbreak scenarios instead of simulations.

Methods

Structural equivalence and homophily. The goal of a node embedding is to represent each node as a low dimensional feature vector while preserving the structure of the graph. That means that the feature vectors of two similar nodes are similar to each other. Node similarity can be defined differently, and strongly depends on the task. The two most commonly studied similarities between nodes in a human contact network are: (1) *homophily*, defined by the shared neighbors of two nodes, and (2) *structural equivalence*, defined by the resemblance in the structural position any two nodes have in the network.

Often, homophily means that connected nodes are similar to each other as they are likely to share similar contacts. As an example, people with similar interests are more likely to meet as they actively seek other people with the same interests. Therefore, a node embedding that preserves *homophily* would be a good choice if the goal was to predict the interests of people in the network. Note that *homophily* can also be modified to consider 2-hop (or n-hop) neighbors or "friends of friends" as similar, which in a social network, might represent an even stronger predictor than a direct connection[9]. However, as further hops are considered, the expected number of neighbors grows exponentially, thereby reducing the uniqueness of each node alongside the usefulness of the similarity metric.

Structural equivalence measures the affinity between two nodes based on the similarity of their position in the network, even though they might not be directly connected. This kind of similarity is often interpreted as a good indicator of the roles people have in society. For example, a manager of a big company would have a different structural position in the contact network than their stay-at-home partner, even though they are in contact frequently.

With *structural equivalence*, two nodes can be considered similar even though they are on different sides of the network, while with *homophily*, similar nodes are always directly next to each other. This makes the two concepts seem opposite, but in human contact networks both may appear at the same time, for example, when a person is similar to their friends, but also to other people in the same role they do not have contact with. For some applications, including disease dynamics prediction, a mixture of the two concepts needs to be preserved and balanced by the embedding algorithm for an accurate representation of society.

`node2vec`. `node2vec` is a node embedding method that allows us to systematically balance *homophily* and *structural equivalence*[10] through a series of random walks. The number of conducted random walks and the length of these walks are hyperparameters of the embedding algorithm. Additionally, `node2vec` has two parameters that control the bias in these exploratory random walks. Depending on these parameters, the random walk is more likely to sample nodes that are either far away (*i.e.,* outward exploration) or close to the start-

ing node (*i.e.,* inward exploration). With inward exploration, the random walk becomes more like the search tree of a breadth first search in the graph, while with outward exploration, it is more similar to a depth first search.

The return parameter `p` controls the probability with which the random walk returns to the last visited node. A high `p` means the random walk is less likely to return, and a low `p` results in nodes that are already in the random walk being sampled more often, yielding even higher similarities between those nodes. To balance this effect, the in-out parameter `q` allows the random walk to favor nodes that lie further away from the starting node. With low `q`, the random walk is biased towards outward lying nodes, specifically, the random walk reaches a greater variety of nodes and those further away can still be considered similar. The authors of the original `node2vec` paper[10] propose that inward exploration preserves *homophily* in the network, while outward exploration preserves *structural equivalence*.

All temporal networks were embedded with `node2vec` using different hyperparameters (see Table S4). Our experiments have shown that the random walk length and the number of random walks do not have much influence on the results, so we fixed those parameters to 10 random walks of size 80. Since `p` and `q` are positive and 1 is the neutral value, we distributed the values for `p` and `q` logarithmically between 0 and 80. We tested 10 different values for each parameter, 5 for inward and 5 for outward exploration. These runs resulted in 100 different possible pairings for `p` and `q`. All of these embeddings were then used to predict the labels gained from 250 SIR-simulations for each network. The prediction algorithm used was logistic regression with l2 regularization.

Datasets. We used 24 datasets, of which 6 are derived from real world data. Five of these datasets were collected between the years 2009 and 2015 at several locations, each collection campaign lasting between 2 days and 2 weeks. In all cases, data were sampled anonymously in order to preserve the identity of the subjects. Furthermore, all subjects or their legal guardians provided explicit **informed consent** in having their data collected and analyzed. The Reality Mining dataset[3] was collected with explicit consent from all subjects by the MIT Human Dynamics Lab, and the use of the set was authorized for such study as long as the privacy of the participants was protected (*i.e.*, no de-anonymization was attempted). The remaining datasets include temporal networks, collected with the consent of the subjects or someone responsible for them[7]. In our analyses, all methods were performed in accordance with the relevant ethical and legal guidelines and regulations.

The empirical data from those datasets were sampled with the same method: a setup of RFID scanners at fixed locations and wearable devices for each subject. A contact was documented whenever two participants were registered by the same set of readers at the same time. In the datasets, all contacts are listed by the IDs of the participants and the corresponding time step with a resolution of 20 seconds[7]. Only one dataset, the Reality Mining dataset, used contacts detected by Bluetooth scans on smartphones. 100 students and faculty members of the MIT Media Laboratory participated in this study, that took place over 9 months in 2004. The scans were conducted every 5 minutes and included timestamps in seconds[3]. Detailed information about the datasets can be found in Table S2. From these data, we derived temporal networks with one node for each participant at each time step and edges as contacts. Data were aggregated to time steps of 10 minutes, a time window that achieves a good balance between the minimal contact duration necessary for a disease to spread and the number of time steps to be analyzed in this study. The other 18 networks were artificially created models of human contact networks. In 9 of them the nodes are connected randomly and the node degree distribution is binomial. In the other 9 networks, the node degree is power-law distributed, which is closer to what has been found in networks derived from real world data[20].

SIR simulation. As reference for how a disease spreads in these networks, a dynamic SIR simulation was conducted. In an SIR simulation, a random node in the network is infected. Then, at each time step, an infected node infects one of its neighbors with probability α (infection rate) and recovers with probability μ (recovery rate). This provides each node at each time step with one of three labels: *susceptible* (S), *infected* (I) or *recovered* (R). In this simplified model of disease progression, "infected" means that a person is contagious, "susceptible" means that a person can become infected and "recovered" means that a person cannot become infected and does not play a role in the disease spreading at this time. In an SIR simulation, the nodes always follow the same disease progression from S to I then R. Note however, that the prediction model is agnostic about the semantics or the order of these labels, and it could just as well predict SIS or SI simulations. We conducted simulations with 5 different parameter sets, which can be found in Table S4.

Network representation. To prepare the dynamic networks for the prediction, the time steps of each network were connected into a static supra-adjacency network. Nodes can be identified by the pair (k, t), where k is the person this node represents and t the current time step. If a node is infected at time step t, it is likely to still be infected at $t + 1$. To use this temporal dependency in the prediction, the time steps of the network are interconnected. This interconnection is done by making node (i, t) always connected to (i, t+1). Furthermore, if there is a contact between person i and person j at time t, there exists an edge from (i, t) to (j, t+1) and from (j, t) to (j, t+1). Additionally, in order to reduce the number of nodes that need to be embedded, only active nodes are considered. These are nodes (i, t) where i had at least one contact at time step t. Inactive nodes are deleted and their incoming edges are rerouted to their next active future self, as previously done by Sato *et al.*[21]. Finally, we converted all 24 datasets into supra-adjacency networks as described. Table S3 shows the different sizes and densities of the real world networks, and Table S1 summarizes those of the artificial ones. In prediction, some of the networks performed better than others, but all showed the described improvements for outward exploration (see Figs. S1, S6, S7 and Table S7).

Evaluation metrics. We primarily chose the f1-score as an evaluation metric for its robustness in evaluating prediction accuracy, by balancing true and false positives or negatives. Since the size of classes can be strongly imbalanced, we look at two different versions to expand the f1-score to multiple labels, namely the *micro f1-score*, which weighs all classes the same, and the *macro f1-score*, that evaluates the score for each class separately and reports the average of all classes. Additionally, we considered different disease-specific metrics that are related to its spreading process:

- End outbreak size, or the number of people that are infected or recovered at the end of the simulation. This number expresses how many people were directly affected by the epidemic.
- Mean prevalence, or the average number of infected people per time step, which indicates the expected outbreak size at any time step.
- Peak prevalence, or the maximum number of people that were infected at any time step. This metric is relevant to estimate the capacity hospitals need to have to provide care for infected people in the worst time of the epidemic outbreak.
- Peak prevalence time, which is the time step when the peak prevalence occurs and is often associated with how aggressively the disease spreads.
- Mean incidence, or the average number of nodes that changed their state from susceptible to infected in one time step. It is indicative of the rate with which the disease spreads.

All of these metrics are evaluated as the difference between simulation and prediction. To meaningfully compare results between different networks, they are given as percentage of participants in the network or percentage of the number of time steps.

Data availibility

The Reality Mining set[3] can be requested from the MIT Human Dynamics Lab at http://realitycommons.media.mit.edu/realitymining4.html. All other contact data[7] can be found at https://doi.org/10.5281/zenodo.2540795. The artificial networks were created with `networkX`[11].

Received: 8 August 2022; Accepted: 13 January 2023
Published online: 17 January 2023

References

1. Askitas, N., Tatsiramos, K. & Verheyden, B. Estimating worldwide effects of non-pharmaceutical interventions on covid-19 incidence and population mobility patterns using a multiple-event study. *Sci. Rep.* **11**(1), 1–13 (2021).
2. Broido, A. D. & Clauset, A. Scale-free networks are rare. *Nat. Commun.* **10**(1), 1017 (2019).
3. Eagle, N. & Pentland, A. Reality mining: Sensing complex social systems. *Personal Ubiquitous Comput.* **10**(4), 255–268 (2006).
4. Fan, C., Lee, R., Yang, Y. & Mostafavi, A. Fine-grained data reveal segregated mobility networks and opportunities for local containment of covid-19. *Sci. Rep.* **11**(1), 16895 (2021).
5. Figueiredo, D. R., Rodrigues Ribeiro, L. F. & Saverese, P. H. P. struc2vec: Learning node representations from structural identity. *CoRR*, arXiv:1704.03165, (2017).
6. Ge, Y. *et al.* Untangling the changing impact of non-pharmaceutical interventions and vaccination on European covid-19 trajectories. *Nat. Commun.* **13**(1), 3106 (2022).
7. Génois, M. & Barrat, A. Can co-location be used as a proxy for face-to-face contacts?. *EPJ Data Sci.* **7**(1), 11 (2018).
8. Gounane, S. *et al.* An adaptive social distancing sir model for covid-19 disease spreading and forecasting. *Epidemiol. Methods* **10**(s1), 20200044 (2021).
9. Granovetter, M. S. The strength of weak ties. *Am. J. Sociol.* **78**(6), 1360–1380 (1973).
10. Grover, A., Leskovec, J. Node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 855–864, New York, Association for Computing Machinery (2016).
11. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring network structure, dynamics, and function using networkx. In editors, *Proceedings of the 7th Python in Science Conference* (ed. Varoquaux, G., Vaught, T., & Millman, J.) 11 – 15, (Pasadena, 2008).
12. Jin, J., Heimann, M., Jin, D. & Koutra, D. Toward understanding and evaluating structural node embeddings. *ACM Trans. Knowl. Discov. Data* **16**(3), 1–32 (2021).
13. Kraemer, M. U. G. *et al.* The effect of human mobility and control measures on the covid-19 epidemic in china. *Science* **368**(6490), 493–497 (2020).
14. Moreno, Y. & Vazquez, A. Disease spreading in structured scale-free networks. *Phys. Condens. Matter* **31**, 10 (2002).
15. Musciotto, F. & Miccichè, S. Effective strategies for targeted attacks to the network of cosa nostra affiliates. *EPJ Data Sci.* **11**(1), 11 (2022).
16. Pinto, E. R., Nepomuceno, E. G. & Campanharo, A. S. Impact of network topology on the spread of infectious diseases. *TEMA* **21**, 95–115 (2020).
17. Qian, X., Sun, L. & Ukkusuri, S. V. Scaling of contact networks for epidemic spreading in urban transit systems. *Sci. Rep.* **11**(1), 4408 (2021).
18. Riad, M. H., Sekamatte, M., Ocom, F., Makumbi, I. & Scoglio, C. M. Risk assessment of ebola virus disease spreading in uganda using a two-layer temporal network. *Sci. Rep.* **9**(1), 16060 (2019).
19. Rossi, R. A., Ahmed, N. K. Role discovery in networks. *CoRR*, arXIv:1405.7134 (2014).
20. Rybski, D., Buldyrev, S. V., Havlin, S., Liljeros, F. & Makse, H. A. Scaling laws of human interaction activity. *Proc. Natl. Acad. Sci.* **106**(31), 12640–12645 (2009).
21. Sato, K., Oka, M., Barrat, A. & Cattuto, C. Predicting partially observed processes on temporal networks by Dynamics-Aware Node Embeddings (DyANE). *EPJ Data Sci.* **10**(1), 22 (2021).
22. Schliski, F., Schlötterer, J., & Granitzer, M. Influence of random walk parametrization on graph embeddings. In *Advances in Information Retrieval* (ed. Jose, J. M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M. J., & Martins, F.) 58–65, (Springer International Publishing, 2020).
23. Song, C., Zehui, Q., Blumm, N. & Barabási, A.-L. Limits of predictability in human mobility. *Science* **327**(5968), 1018–1021 (2010).

24. Tantrakarnapa, K., Bhopdhornangkul, B. & Nakhaapakorn, K. Influencing factors of covid-19 spreading: A case study of thailand. *J. Public Health* **30**(3), 621–627 (2022).
25. Tian, L. *et al.* Harnessing peak transmission around symptom onset for non-pharmaceutical intervention and containment of the covid-19 pandemic. *Nat. Commun.* **12**(1), 1147 (2021).
26. Tu, K., Cui, P., Wang, X., Yu, P. S. & Zhu, W. Deep recursive network embedding with regular equivalence. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '18, 2357–2366 (Association for Computing Machinery, New York, 2018).
27. Wang, L., Huang, C., Ma, W., Lu, Y. & Vosoughi, S. Embedding node structural role identity using stress majorization. *CoRR*, arXiv:2109.07023 (2021).
28. Xiong, C., Songhua, H., Yang, M., Luo, W. & Zhang, L. Mobile device data reveal the dynamics in a positive relationship between human mobility and covid-19 infections. *Proc. Natl. Acad. Sci.* **117**(44), 27087–27089 (2020).
29. Zhang, B., Pavlou, P. A. & Krishnan, R. On direct versus indirect peer influence in large social networks. *Inf. Syst. Res.* **29**(2), 292–314 (2018).

Author contributions

P.K. and L.T. designed the research, P.K. conducted the experiments and prepared all figures and tables, P.K. wrote the first draft of the text, and L.T. finalized the editing of the text and supervised the work. All authors reviewed the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-28126-w.

Correspondence and requests for materials should be addressed to L.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023

Non-Evaluation Relevant Publication - Publication 8

Title: Ethical and Privacy Considerations with Location Based Data Research
Authors: **Leonardo Tonetto** (TUM), Pauline Kister (TUM), Nitinder Mohan (TUM), Jörg Ott (TUM)
Journal: ArXiv
Publishing date: February 11, 2024
Reference: [196]

This preprint is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0)

©2025 Leonardo Tonetto, Pauline Kister, Nitinder Mohan, Jörg Ott. This work was first made available on arXiv at [https://arxiv.org/abs/2403.05558]

Ethical and Privacy Considerations with Location Based Data Research

Leonardo Tonetto
Technical University of Munich
Garching bei München, Germany

Pauline Kister
Technical University of Munich
Garching bei München, Germany

Nitinder Mohan
Technical University of Munich
Garching bei München, Germany
mohan@in.tum.de

Jörg Ott
Technical University of Munich
Garching bei München, Germany
ott@in.tum.de

ABSTRACT

Networking research, especially focusing on human mobility, has evolved significantly in the last two decades and now relies on collection and analyzing larger datasets. The increasing sizes of datasets are enabled by larger automated efforts to collect data as well as by scalable methods to analyze and unveil insights, which was not possible many years ago. However, this fast expansion and innovation in human-centric research often comes at a cost of privacy or ethics. In this work, we review a vast corpus of scientific work on human mobility and how ethics and privacy were considered. We reviewed a total of 118 papers, including 149 datasets on individual mobility. We demonstrate that these ever growing collections, while enabling new and insightful studies, have not all consistently followed a pre-defined set of guidelines regarding acceptable practices in data governance as well as how their research was communicated. We conclude with a series of discussions on how data, privacy and ethics could be dealt within our community.

KEYWORDS

human mobility, ethics, privacy

1 INTRODUCTION

Understanding human mobility has been a quest of scientific research for many years, but it was not until the popularization of smart phones and its embedded sensors that we could deeply study it. The field itself has grown tremendously in the past two decades. From survey-based studies with limited number of subjects, sensing human movements through mobile systems enabled large scale studies benefiting the design of better communication protocols, urban policies and containment of infectious diseases. As the research problems and scope in the field evolved and grew, so did the data collection methods utilizing new sensing technologies, multiple experiment sources, number of involved subjects, etc. As such, the human mobility datasets have become larger and more diverse, more robust and complex methods and models were devised to study these sets, and a vast amount of work has revealed potential privacy implications with such data (*e.g.*, de-anonymization methods [46, 65, 164]).

To handle the growing ethical concerns from research studies involving human subjects, outlets for scientific articles are increasingly requiring ethical statements from authors at the time of submission itself. For example, venues associated with ACM, such as SIGCOMM (*e.g.*, IMC, HotNets, CoNEXT) and SIGMOBILE (*e.g.* MobiCom, MobiSys, MobiHoc) as well as workshops on the topic require authors to refer to their principles and code of ethics [48, 109] and submit a statement accompanying the article explicitly mentioning any ethical concerns raised by the study and steps undertaken by authors to mitigate them [2]. However, most of these policies are still evolving and ethical issues from published scientific studies and datasets have still not been eradicated [113, 152]. We believe that the primary contributor to this is the metamorphosis in the definition of "ethics" in research as the views around data ownership in collection evolve with technological advancements. However, untimely discovery of ethical issues within research not only jeopardizes the privacy of involved subjects but can cause embarrassment to involved researchers through dataset redactions, erratas (sometimes retractions [118]) along with widespread clarifications.

Despite serious implications, no recent study has reviewed and discussed how research articles, especially focusing on network research, used personal data of its subjects whilst dealing with privacy and ethics. In this paper, we plug this gap in research and focus on research on individual mobility data – as those have been extensively used in recent networking research (*e.g.*, [75, 103, 168]) and arguably are a source of data the general public (outside of the research community) relates to most (*e.g.*, [154]). In this article, we do not aim at reviewing the latest methods for studying human mobility, but rather we reexamine how individual location data was used in those studies, in light of recent developments in data regulations and ethics, along with the different properties of the data used. To this purpose, we discuss **(1)** the various

types of datasets used for mobility research, **(2)** recent developments in data de-/anonymization that may shape public opinion on ethical concerns involving different data types, **(3)** the current state for regulations in certain countries for the use of individual location data, **(4)** how human mobility papers have treated this matter over the years, and finally **(5)** we discuss approaches future human mobility research could follow to establish a sustainable environment for both researchers and subjects.

To the best of our knowledge, ours is the first work to take a broader, multi-faceted look at the evolving state of human mobility research through ethical lenses. We believe that the observations and discussions raised in this work can extend to other fields of research, wherever personal information of human subjects is involved. As we point out later in this paper, we believe that more could be done to unify and standardize best practices on data governance as well as how research results should be communicated. The relevance of this topic lies in (1) protecting the privacy of those subjects being studied, and (2) preserving the trust of the general public with such research in a preventive manner, allowing for even more studies to be published.

2 RELATED WORK

In recent years, a series of scientific papers have discussed different aspects of ethics in networking research. While Partridge *et al.* [113] advocated for an "*ethical considerations*" section in all network measurements papers, the work by van der Ham *et al.* [159] and Thomas *et al.* [152] reviewed the use of datasets in network measurements research. The work by van der Ham *et al.* [159] discussed the ethical implications of four case studies where publicly available data contained more information than expected, compromising the privacy of their subjects. The work by Thomas *et al.* [152] reviewed the ethics of using datasets of illicit origins for research. Their work discussed the limitations regarding the lack of informed consent in such studies whenever humans are part of the leaked data. Furthermore, they present a series of case studies where leaked data were used in security related scientific papers. The authors present and rebut some commonly used arguments to justify the use of such data, and conclude that both researchers and scientific outlets should take a more responsible and ethical approach towards researching on stolen data. On the same topic, Ienca *et al.* [82] reviewed the use of hacked data in machine learning papers, suggesting pondering benefits and risks of each dataset while being clear about the means through which data were obtained and how researchers dealt with personal information, if present. From a different angle, we discuss how datasets legally obtained were used, while containing personal data and how these were communicated in their published manuscript.

The ethical implications of re-using data as well as reproducibility of results have also been addressed. Boté *et al.* [31] discussed the ethical and privacy problems associated with the re-usability of research data, arguing for its benefits regarding reproducibility and further research results from an already existing set. Furthermore, reviewing a series of Internet measurement papers, a recent study by Demir *et al.* [47] reviewed the reproducibility of methods used in those papers, along with observations about the content of these papers. The authors observe that almost two-thirds of the analyzed papers (N=117) do not provide an ethics section, in spite of mostly focusing on security and privacy work. We make similar observations for papers using human mobility data, which is arguably a topic likely to reach non-scientific audiences. Additionally, we discuss other aspects related to privacy, such as IRB checks, consent

Furthermore, while Iqbal *et al.* [3] reviewed the progress and citations of ACM SIGCOMM papers over recent decades, an early study by Kurkowski *et al.* [93] discussed how various methods for simulating mobility were used in the early 2000s. In this work, we include a survey covering over 20 years of research and how mobility datasets have evolved.

3 HUMAN MOBILITY DATASETS

Individual human mobility datasets may be collected from multiple sources, but are fundamentally described in the same way. They capture the historical presence of a person (or a handheld smart device) at a given point in space and time. Next we present a logical separation of the main classes of mobility data, with its main characteristics summarized on Table 1.

3.1 Communication Infrastructure based

This source of data relies on existing communication infrastructure to sense the presence of a subject. This presence is inferred by records (or logs) created when a subject's mobile device communicates with a point of access, which could be of a Wi-Fi network or of a mobile cellular network [29]. While Wi-Fi setups are often limited to confined areas, such as companies or universities [11, 35], they offer room-level accuracy for their locations. Mobile cellular networks on the other hand may cover entire countries but have their accuracy between 100s meters to 10s kilometers [70] with call detail records (CDR). In both cases, the availability of location records may be a function of the activity of the user, such as when an incoming/outgoing phone call happens or a device associates to an access point upon arriving to a new area. Additionally, certain network settings allow location records to be captured whenever *any* data or even signaling events happen between the network and a subject's device (*e.g.*, eXtended Detail Records (XDR) and control panel records

Table 1: Summary of data types used for human mobility studies.

Class	Source	Location Accuracy	Sample Trigger	Area Covered
Communication Infrastructure	Telco	100 m - 10 km [111]	CDR: calls XDR: calls+data CPR: cell signalling	city, country
	CCR	room, building [51]	user purchase	city, country
	Wi-Fi	room, building [35]	de-/association	campus, city
			+ active scanning	any
Experiments	GPS sensor	10 m - 100 m [140]	phone setting	any
	Bluetooth	10 m - room [140]	phone setting	any
Internet Based Services	OSN	10 m - 100 m [88]	content post	any
	Web	100 m - 10,000 km [73]	page/app interaction	any
Public Transport Infrastructure	SmartCard	0 m - 10 m [142]	get on/off transport	city

(CPR) [111]). For such, subjects often agree on *terms and services* for the use of the offered infrastructure, and the collection of location data can be seen as a byproduct of this interaction. Alternatively, Credit Card Records (CCR) can also provide rich information about its user's whereabouts whenever a card is used for payment at a physical store [51].

3.2 Designed Experiments

Another class of human mobility data source originates from designed experiments, in which a phone app or preconfigured devices are distributed among subjects (*e.g.*, [59, 140]). These types of effort often provide the highest level of flexibility and uniformity in how data are sampled, at times also providing data from other non-location sensors, such as accelerometers. These extra readings enable a higher accuracy in segmenting events, such as the duration of stops. However, as these studies are based on recruiting people, and given its costly setup and lack of secondary benefits for subjects (such as Internet access through a Wi-Fi access point), cohorts are limited in their size when compared to communication infrastructure based efforts [140]. Location data collected through these studies often include continuous GPS coordinates, Bluetooth (BT) and Wi-Fi scans of nearby devices.

3.3 Internet Based Services

Another infrastructure provider that captures location data are those enabling the exchange of information on the Internet. Web services, such as online social networks (OSN) [88], search engines [170] and map services [170] provide online users with services while also logging their physical location, for example by means of geotagged posts or geolocating devices based on their IP address [32, 73]. Unlike the sources previously discussed, the sampling of location records by web services is highly dependent on how often a subject interacts with those services, leading to a skewed availability of data per device. Alternatively, location data may be captured in the background but only if allowed by the user (*e.g.*, [16, 71]). Additionally, these web services can also capture extra features, such as the content of what is being searched or posted as well as the social graph of their users, which can be used to further enrich any analysis being made. Similar to the communication infrastructure, subjects agree to *terms and conditions* for that service that states their location data will be logged and may (or may not) be used for further studies.

3.4 Public Transport Infrastructure

Smart Cards have replaced old paper-based ticketing system in most modern public transportation systems in large metropolitan areas [142]. They provide an integrated and automated way for passengers to pay for transport rides as well as manage different pricing schemes (*e.g.*, senior citizens or students discounts). Users are required to present their smart cards before starting a ride, for example when entering a subway station or boarding a bus, and in some cases the same is expected when alighting. In this way, the system records the timestamps of discrete location points a subject has been. As human mobility tends to produce repeatable patterns [136], location data from public transport tends to be consistent and homogeneously through time, at least until a global pandemic changes how people move. Similar to all

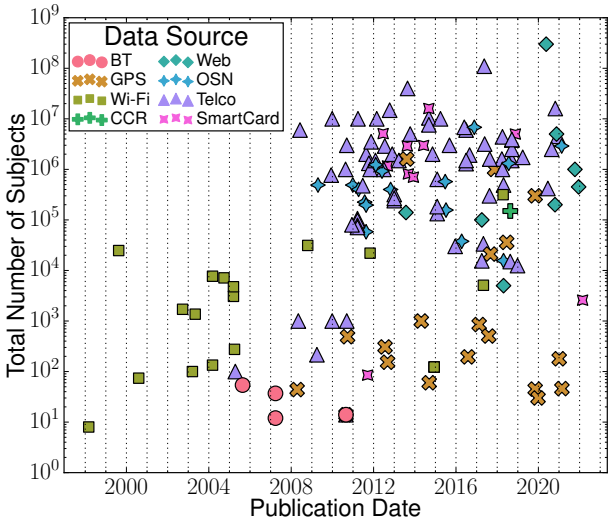


Figure 1: Size of datasets per publication date.

infrastructure-based sources, subjects agree to the *terms and services* of using a smart card for their corresponding transport system. Additionally, smart cards have also been used to trace the behavior of students in a university campus, logging various activities and services used by students [98]. Other examples of mobility data gathered from public transport are shared bikes [1], taxis [146].

Given the various data sources discussed above, we now turn to a set of observations drawn on the scales of the papers surveyed and their respective datasets.

4 MEASUREMENT METHODOLOGY

4.1 Literature Selection

Clarity and *justifiable choices* are of paramount importance when personal human data are used as an input to a research study. To understand how individual mobility data has been utilized in research till date, we conduct a survey on most relevant outlets for computer science with special focus on networking research venues (*e.g.*, IMC, Mobicom, INFOCOM) and journals involving general sciences (*e.g.*, PNAS, Nature, Science). Within this dataset, we filter for works that used human mobility dataset through keywords such as *human mobility* and *individual location data*, as well as including their respective relevant citations. We further filtered studies where authors clearly stated that they had access to some form of user identifier (anonymized or not) and multiple spatial points per subject (*e.g.*, geographical coordinates, points of interest) in the description. To limit the scope of this review, we explicitly excluded research from *Transportation* and *Geographic Information Systems* fields. We also did not consider any *car/vehicle* dataset, as those often involved taxi drivers or cars with multiple owners (*i.e.*, encompassing more than one subject).

We performed this survey across the past 24 years (1998-2022) resulting in **118** research articles. Of these 118 papers we surveyed, we identified a total of **149** datasets with varying properties based on the description given in their respective papers. We categorized the papers in *six* different buckets based on their method for estimating locations of involved subjects, specifically bluetooth (*BT*), *GPS*, *WiFi*, credit card records (*CCR*), web browser (*web*), online social networks (*OSN*), telecommunication networks (*telco*) and public transport smartcards (*smartcard*). We observed that, thanks to the advancements in data collection methods, the collected sizes of mobility data has grown exponentially over the years (see figure 1). Furthermore, we also observed that recent studies on the subject tend to incorporate more than one data sources to collect information of its subjects, as discussed in the previous section – further expanding the involved dataset sizes, both in number of participants and the amount of collected data.

4.2 Scale of Selected Studies

With a growing deployment of Wi-Fi networks at universities and company facilities, logs from these setups have allowed researchers to obtain reliable mobility data from hundreds and even thousands of residents, including from their dormitories. While several mobility models were built with such data (*e.g.*, [22]), these were mostly limited to workdays and covered only a small geographical area. To overcome that limitation, the late 2000s and early 2010s saw the emergence of a wide variety of mobility data sources, with even larger pool of subjects and, in a few cases, even covering multiple countries. From our analysis we also observed that while the number of subjects increased over time and larger data collections periods were used, although those did not correlate with the cohort sizes. Additionally, Figure 2 depicts the distribution of dataset sizes per data source, further highlighting the disparities in cohort sizes between studies that often provide informed consent or not, as will be discussed later in this article. Unsurprisingly, sets covering larger areas also tend to include more users, as shown in Figure 3. However as multiple countries are only included in OSN and Web collections (see their limitations above) those do not necessarily have the largest groups.

We also find that a majority of articles tend to not properly describe the properties of their dataset well within their text. Surprisingly, almost half of the datasets (81, or 54.4%) within the selected articles did not report the number of records (or table rows) used in the study. For those which reported those numbers, we also observed a similar growth trend for overall

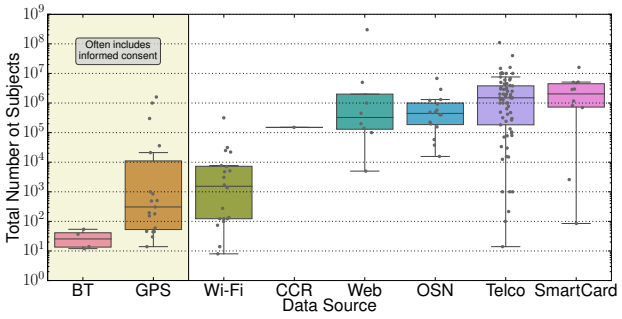


Figure 2: Size of datasets per source.

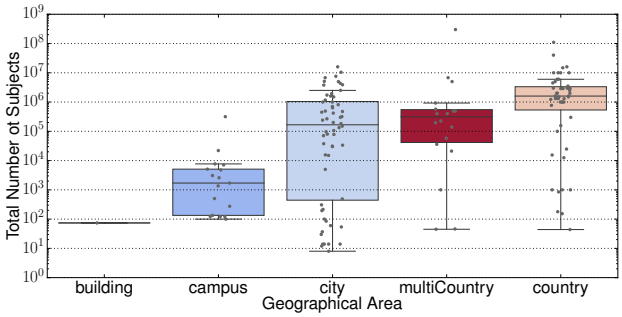


Figure 3: Size of datasets per geographical area/type coverage.

number of rows in datasets over the years. Only one in eight of the total datasets (19, or 12.8%) reported the number of rows used (*i.e.*, after initial filtering), and for those which reported values, on average half (50.4%) of the total rows were discarded. This filtering is typically done to reduce temporal sparsity without significantly reducing temporal resolution. Similarly for the number of subjects available for each study after filtering, one in four (43, or 28.9%) report total and filtered values where on average nearly two thirds (62.1%) of subjects are discarded. This filtering, while necessary for a homogenized set of subjects over time, may also introduce biases to the results, which should also be discussed. In Section 7 we discuss how the papers we surveyed often had access to more information about their subjects than it was actually needed for their studies, such as *home* location and their social graph.

Our analysis on the scale of mobility datasets provides context to the increasing magnitude in the number of subjects and rows as well as in geographical extent of measurement studies in the field over the years. While the scales of datasets used for such studies has grown significantly, so has the general public's awareness of possible privacy issues related to those data [14]. Complementary to this awareness, as new research on anonymization methods became popular, new

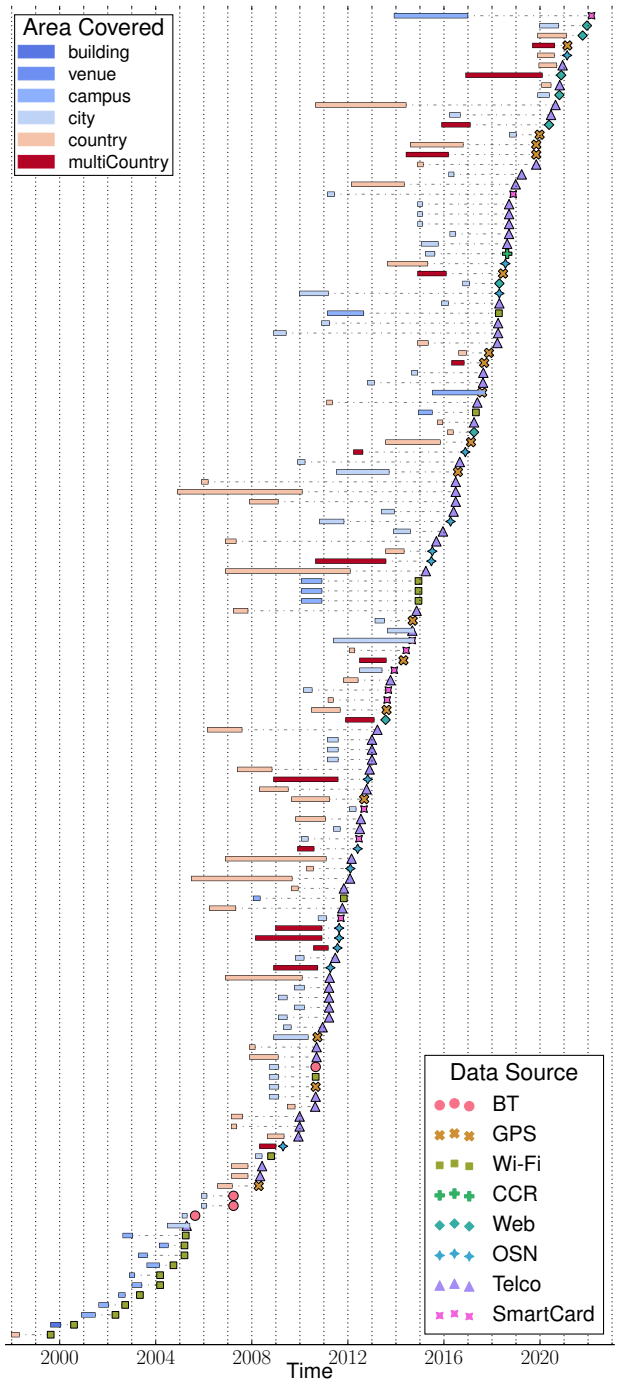


Figure 4: Data source, sampling duration and area type per dataset. Entries sorted by date their papers were published.

de-anonymization approaches have followed suit. Next, we

review the literature on such methods as they support our understanding of how future research on human mobility ought to follow.

5 THE PRIVATE DATA CONUNDRUM

In 2006, AOL shared a sample of historical queries from over 650,000 users for research purposes. The identities of these users were altered in order to protect their privacy, but shortly after its release a series of users were *de-anonymized* making the headlines of newpapers (*e.g.*, [21, 130]). Furthermore, even data not released on purpose may be leaked through security breaches (*e.g.*, [24]). Consequently, these events populate public opinion about data security and privacy, leading to new or modified legislation as well as ethical guidelines. These new directions finally shape recent network research by limiting what can be done when personal data are used, or at least it should.

To tackle these constraints when dealing with personal data, researchers have paid significant attention to methods for both anonymizing private personal records, simultaneously to de-anonymize it. These works often aim at raising awareness of possible weaknesses in identity protection, and how to provide minimal guarantees to one's privacy, respectively.

Even though we do not analyze the full life-cycle of datasets, in several occasions sets were either made public or at least used in further studies carried by the original research group. On that topic, Allman and Paxson addressed the ethical concerns of releasing and using data, including privacy protection [13]. They emphasize that raw datasets containing identifiable information should not be shared publicly, and research should prioritize preserving the privacy of individuals. Additionally, the authors advise against trusting anonymization approaches as deanonymization counter-attacks quickly become available.

We now provide a brief review of some of such methods as those are often cited, as we observe, as reasons for an unconditional care whenever individual data are used.

5.1 Location-based data de-anonymization techniques

When sharing individual data, personal attributes which can *directly* identify any subject (*e.g.*, name or phone number) may be changed to *pseudonyms* in order to protect their privacy. These steps, however, yield only partial privacy protection as highlighted by a plethora of studies on de-anonymization attacks [62, 66]. Following the taxonomy of Fung *et al.* [66] and Fiore *et al.* [62], we review some of these attacks which we later refer to when discussing individual location data in mobility research. The relevance of these studies to our current work lies in the observation that all papers we surveyed state that user identifiers were *anonymized*, and in some cases that being the only privacy related content in the manuscript.

On a study on attacks through *record linkage*, in which an attacker has a database containing (quasi-)similar location records from their victims, De Montjoye *et al.* [46] showed how any two location records could de-anonymize 50% of subjects and how four locations could increase this trajectory uniqueness up to 95%, from a set of 1.5 million persons. However, a comprehensive study by Wang *et al.* [164] showed that former methods actually underperformed when applied to real-world large scale data, reporting a de-anonymization hit rate lower than 20% in all cases. Wang *et al.* then finally proposed a Markov-based model that achieves a 40% success rate in linking records from heterogeneous databases.

Another relevant surface of attack is through a side-channel, such as a victim's profile. By using both *home* and *work* information of victims, Freudiger *et al.* [65] show how an adversary can uniquely identify individuals in large datasets with up to 90% accuracy. Furthermore, knowledge on the whereabouts of one's friends, or her *social graph*, may reveal relevant information about her current location. Using data from online social networks, Sadilek *et al.* [123] proposed an unsupervised method, achieving 57% and 77% accuracy in predicting a victim's location given information from two and nine of her friends, respectively. Similarly with a supervised learning approach, the authors obtained accuracies of 77% and 84.3%, respectively. Using a different approach and a set of smaller but denser networks, Srivatsa *et al.* [139] reported that up to 80% of subjects' identity could be recovered by matching social graphs links with increased physical proximity.

Taken together, these attacks are vivid examples of how individual location data can uniquely identify a person and what some vectors for de-anonymization are. These threats may shape not only the contents of academic research but also the general public view on the harm such data impose to their individual privacy. To counteract these perils and uncertainties two alternatives are often presented: (1) stronger anonymization methods, and (2) ethical guidelines and regulations aimed at protecting a person's right for anonymity. We further discuss both of these point, next.

5.2 Stronger anonymization for location-based data

Given the threats to an individual's privacy brought by the de-anonymization attacks discussed above, a growing literature on anonymization methods has produced relevant results. Following the taxonomy of Fiore *et al.* [62], anonymization

techniques follow three main principles. (1) *Indistinguishability*, in which the records of a subject must be indistinguishable from those of the same size-limited anonymity set, essentially removing the unicity from one's trajectory. A commonly used formal approach for indistinguishability is *k-anonymity* [143]. (2) *Uninformativeness*, in which an adversary's knowledge about a victim must remain unchanged after accessing a dataset the victim is part of. This principle is generally achieved through differential privacy [57]. (3) *Mitigation*, in which preventive measures are taken to assure one's data privacy, without a well-define principle. Such principle is addressed with alterations in the original location data, aimed at discarding certain unique aspects of one's trajectories, however without providing formal guarantees.

On *indistinguishability*, ensuring *k-anonymity* has taken a multitude of solutions, such as: (a) *spatiotemporal generalization*, in which spatial and temporal resolutions are reduced until traces from members of an anonymity set are indistinguishable. Achieving this goal, however, quickly reduces the quality of an anonymized location dataset as more trajectories are aggregated, as revealed by De Montjoye *et al.* [46]. (b) *Suppression*, in which points are removed from trajectories until their unicity is lost [149]. (c) *Spatial uncertainty*, in which *k-anonymity* is relaxed to include entire trajectories that fall within a defined threshold, as demonstrated by Abul *et al.* [4]. It is important to note that *k-anonymity* has its own shortcomings, addressed by extensions such as *l diversity* that enforces an extra diversity criteria per group [104].

On *uninformativeness*, proposed solutions formally follow differential privacy in which consecutive similar queries to a dataset should return similar results, even if the trajectories of a subject have been added or removed from the set. This has been shown, for example, by Shokri *et al.* [132] who adapted differential privacy for location data. Other examples of differential privacy include altering the representation of the trajectory data using different data structures [37] as well as aggregating observations as probability distributions [74].

On *mitigation*, proposed solutions do not provide formal privacy guarantees and also often reduce the utility of the original location data. This is achieved through: (a) *obfuscation*, in which noise is added to location data (*e.g.*, [55]). (b) *Cloaking*, in which spatial or temporal resolutions are reduced (*e.g.*, [122]). (c) *Segmentation*, in which new pseudonyms are used for multiple segments that are artificially added to the data (*e.g.*, Song *et al.* [138]). (d) *Swapping*, in which pseudonyms are randomly swapped between different trajectories (*e.g.*, [126]).

The observations in these studies remind us that replacing real identifiers by pseudonyms is not enough to protect one's privacy. A myriad and ever growing literature on de-anonymization approaches highlights not only the risks of dealing with personal data, but also the importance of stronger anonymization techniques. These protective methods, however, often degrade information present only in the original datasets. This inevitable drawback affects what studies are capable of doing, by eliminating important nuances, as well as whether or not data can be shared, and in which format.

6 ETHICAL GUIDELINES AND REGULATIONS OUTLOOK

Given all the risks to one's privacy regarding location data, ethical guidelines as well as regulatory bodies have adapted their take on these topics. These efforts are of paramount importance as they aim at protecting subjects in circumstances where only hindsight can assure privacy, even if only partially. Ethical considerations are relevant in all stages of human mobility research, from the design of research questions, the analysis of data, the articles being reviewed and published as well as to the dissemination of the work that is either being done or that is already completed [64].

To safeguard individual's privacy, the European Union General Data Protection Regulation (GDPR) noticeably covers exclusively personal data. In the GDPR, personal data are defined with respect to the identifiability of an individual based on the data being regarded. However, if data are not personal or have been sufficiently anonymized they are no longer protected by the GDPR laws [5]. As we previously discussed, there is a large body of research on how most methods of anonymization are flawed and do not guarantee lack of unicity. Therefore, human mobility research for often dealing with personal location data has to ensure not only laws are met, but that research is conducted and presented following up-to-date ethical guidelines. This would ensure public *trust* and the continuity of this research area.

With a similar purpose, the California Consumer Privacy Act (CCPA) [108] establishes a set of rules to protect the privacy of online users, with special regards to the sale of personal data. A fundamental difference between the CCPA and the GDPR concerns the origin of the data covered by each law. While the former covers data provided by the consumer the latter makes no distinction, as long as it is identifiable.

An inexorable aspect of privacy, both ethically and in various law instances, is the need for consent. From a subject's point of view, individuals have reported being more willing to consent with the use of their data if the proposed research had clear benefits, either personally or for society [156]. That is, if one assumes a utilitarian view, where benefits may justify potential risks, subjects are more likely to agree in taking part in a study if either them or society may benefit from the results of the study. One issue here, we argue, is that unlike medical research, human mobility studies from computer

or social scientists are less tangible. Therefore, drawing a line to what is or is not ethical becomes harder. Nevertheless, mobility research can still assure subjects when data are being collected, what analyses are being done, and what the overall impact might be as those elements define public opinion on whether or not their data should be used [19, 134]. We note, however, that ensuring this transparency between researchers and subjects is often referred in the surveyed papers as one of the main reasons *no data* can be shared. As a result, the *reproducibility* of results is impaired, which we further discuss in Section 8.

Another important facet of this complex problem is the ethical tradition of researchers and subjects' culture in different countries [64]. Traditionally, US and UK scientists are more *utilitarian* in which the value of the research being done justifies the potential ethical risks associated, whereas their European counterparts favor a more *deontological* view, *i.e.*, it favors moral values as well as rules before weighing and considering the outcome. Regarding subjects that in some studies may be from several countries, such as those using OSN data, their personal and cultural understanding and assumptions on privacy are *ambiguous, contested and changing* [64]. That is, given this vast combinatorial of cultural understandings and expectations regarding privacy and its associated risks, a cautious and preventive approach is undoubtedly beneficial in helping guide future scientific (and possibly commercial) efforts using personal location data.

There are, however, a series of steps human mobility research can follow in order to safeguard subjects' privacy as well as scientific and general public's trust on what is being studied. Even though these actions often require additional intricate effort from researchers, we observe an increase in its adoption by publication outlets. These include: (a) reviews from ethical review boards (or institutional review board, ERB/IRB), with these reviews/assessments being required to be independently verifiable (*e.g.*, with a protocol number); (b) including comprehensive ethical statements, where the possible *vulnerability* of and *harm* to subjects is discussed, and what might have been the expectations of users of a studied service when interacting with it; (c) whether or not informed consent was given to the specific study being conducted, beyond the *terms and services* of that platform providing the location data; and finally overall; and (d) what was done to minimize *risk* for subjects, given all that has been discussed thus far. In the next section we review how, in the past, research papers have addressed some of the outstanding challenges human mobility faces regarding ethics, followed by a discussion on how we believe this should be done in the future.

7 ETHICAL CONSIDERATIONS IN THE LITERATURE

In this section we discuss how human mobility research papers dealt with some of the ethical aspects presented in the previous section. We present numbers based on a review of 118 papers from CS and general sciences, published from 1998 and 2022, which used individual location data, totalling 149 datasets (*i.e.*, papers could contain multiple sets). In all cases, we consider only the first publication to include a given dataset (see § 4), and a basic of description of all papers is presented in Table 2.

We begin by making a distinction between datasets collected *deliberately* for a study and those which were shared with researchers through an *agreement*. From all datasets, one in four (37, or 24.8%) were collected *deliberately*, as an integral part of the study. Additionally, sets covering larger areas, or that were collected more recently, or sets with larger subject counts showed a lower likelihood of a deliberate collection effort. This could, in part, be explained by a predominant use of CDR and OSN/Web based data in recent years, as depicted in Figure 1, as those sources often capture location data for basic operations, such as billing, and which are then repurposed for mobility research.

Data sharing remains the exception, understandably. While data sharing has grown in importance for reproducing and validating existing results or fostering new research, privacy concerns and NDA's often prevent any data being shared. From all sets, one in ten (16, or 10.7%) were shared in either their original form or modified along with the paper (*e.g.*, [91, 101]). However, we note that, in some cases, authors provide the contact of the data provider stating that those sets could still be retrieved upon agreement.

In some cases, papers took deliberate actions to protect subjects privacy, beyond anonymizing user identifiers one single time. For that, we identified one in 15 (10, or 6.7%) of sets received extra steps, with no clear distinction between CS and general sciences. Additionally, we note that such measures became more common in more recent papers. Examples of such mitigating measures include *spatial aggregation* (*e.g.*, [9, 26, 92, 100]) where records per individual are still available but the spatial resolution is reduced, similarly with *temporal aggregation* (*e.g.*, [103]) where records are grouped in large time slices, and *periodical changes in user identifiers* (*e.g.*, [75]). It is worth noting that only some papers state what might be clear, that "*obeying data privacy regulations is important when analyzing mobile phone usage data*", and after explaining what was done conclude that "*thereby the local regulations have been met and the recommendations of the GSMA, the alliance of mobile phone providers have been followed.*" [75]

Table 2: Papers surveyed. *Year* published, if *Deliberate* privacy was considered, if an ethical *Statement* was present, if an *IRB* approval was present, if subjects gave *Consent* to the exact terms of the study, and which data *Sources* were used. Sources: W - Wi-Fi; T - Telco; B - Bluetooth; O - Online Social Networks; S - Smartcards; E - Web Services; C - Credit Card Records. (● - Yes; ○ - No)

Paper	Year	Deliberate	Statement	IRB	Consent	Sources
[94]	1998	○	○	○	○	W
[144]	1999	○	○	○	○	W
[145]	2000	○	○	●	○	W
[81]	2002	○	○	○	○	W
[90]	2002	○	○	○	○	W
[150]	2003	○	○	○	○	W
[20]	2003	○	○	○	○	W
[76]	2004	○	○	○	○	W
[39]	2004	●	○	○	○	W
[129]	2004	●	○	○	○	W
[105]	2005	○	○	○	○	W
[59]	2005	○	●	○	●	T
[158]	2005	○	○	○	○	W
[80]	2005	●	○	○	●	B
[35]	2007	●	○	○	●	B
[121]	2008	○	○	○	○	G
[6]	2008	○	○	○	○	W
[70]	2008	○	○	●	○	T
[136]	2009	○	○	○	○	T
[147]	2009	○	○	○	○	T
[58]	2009	●	○	○	●	T
[43]	2009	○	○	○	○	O
[44]	2010	○	○	○	●	G
[42]	2010	○	○	○	○	O
[119]	2010	○	○	○	○	T
[135]	2010	○	○	○	○	T
[85]	2010	○	○	○	○	T, G, W, B
[115]	2010	○	○	○	○	T
[89]	2011	○	○	○	○	T
[17]	2011	○	○	○	○	T
[26]	2011	●	●	○	○	T
[95]	2011	●	●	○	●	S
[128]	2011	○	○	○	○	O
[83]	2011	●	●	○	○	T
[116]	2011	○	○	○	○	T
[114]	2011	○	○	○	○	T
[106]	2011	○	○	○	○	W
[40]	2011	○	○	○	○	O, T
[38]	2011	○	○	○	○	O
[123]	2012	○	●	○	○	O
[96]	2012	○	○	○	○	S
[45]	2012	○	○	○	○	T
[30]	2012	○	●	○	○	T
[107]	2012	○	○	○	○	O
[120]	2012	○	○	○	○	T
[54]	2012	●	○	○	●	G
[12]	2012	○	○	○	○	O
[169]	2012	○	○	○	○	T
[124]	2012	○	○	○	●	G
[171]	2012	○	○	○	○	S
[133]	2012	○	○	○	○	T
[99]	2012	○	○	○	○	T
[27]	2013	○	○	○	○	S
[137]	2013	○	○	○	○	G
[100]	2013	●	○	○	○	T
[46]	2013	○	○	○	○	T
[142]	2013	○	○	○	○	S
[25]	2013	○	●	○	○	T
[117]	2013	○	○	○	○	T
[170]	2013	○	○	○	○	E
[179]	2013	○	○	○	○	S
[162]	2014	●	●	●	○	W
[167]	2014	●	●	●	●	G
[140]	2014	●	●	●	●	G
[49]	2014	○	○	○	○	T
[181]	2014	○	●	○	○	T, S
[141]	2014	○	○	○	○	S
[163]	2015	○	○	○	○	T
[112]	2015	○	○	○	○	T
[97]	2015	○	○	○	○	O
[23]	2015	○	○	○	○	T
[155]	2015	○	○	○	○	T
[88]	2015	○	●	●	○	O
[87]	2016	○	○	○	○	T
[50]	2016	○	○	●	○	T
[160]	2016	●	●	●	●	G
[110]	2016	○	○	○	○	T
[77]	2016	○	○	○	○	O
[151]	2016	○	○	○	○	O
[153]	2017	●	○	○	○	T
[102]	2017	○	●	●	○	T
[161]	2017	●	○	○	●	G
[173]	2017	○	○	○	○	E, T
[180]	2017	○	○	○	○	G
[15]	2017	●	○	●	●	G
[10]	2017	○	○	○	○	T
[86]	2017	○	○	○	○	T
[34]	2017	○	○	○	○	W
[33]	2017	○	○	○	○	T
[8]	2017	●	●	●	●	G
[9]	2018	●	●	●	●	G
[168]	2018	●	○	○	○	T
[174]	2018	○	○	○	○	T
[166]	2018	○	○	○	○	O
[36]	2018	○	○	○	○	T
[60]	2018	○	○	○	○	T
[11]	2018	○	○	○	○	W
[125]	2018	○	○	○	○	T
[78]	2018	○	○	○	○	S
[61]	2018	○	○	○	○	O, E, T
[51]	2018	○	○	○	○	C, T
[148]	2019	●	○	○	○	G, T
[41]	2019	○	○	○	○	G
[157]	2019	●	○	○	○	G
[165]	2019	●	●	●	○	T
[75]	2020	●	●	○	○	T
[103]	2020	●	●	●	○	T
[69]	2020	○	○	○	○	T
[92]	2020	●	●	●	○	E
[7]	2020	○	●	●	○	E
[67]	2020	○	●	●	○	T
[175]	2020	○	○	○	○	E
[176]	2021	●	●	●	●	E
[131]	2021	○	●	○	○	E
[84]	2021	●	○	○	●	G
[79]	2021	○	○	○	○	O
[172]	2021	●	○	●	●	G
[98]	2022	○	○	○	○	S

As discussed previously, beyond the use of user identifiers, subjects can be de-anonymized through other bits of information, such as the places they visit, the schedule they follow, and their social network. Therefore, approaches can be taken to minimize the availability of such information, which in many cases still allows the same research questions to be answer. One of these approaches is a periodic change of anonymized identifiers per subject (*e.g.*, daily), which was observed in only one in 16 (9, or 6.0%) of the datasets we observed. Knowing the home of a subject (or approximate location) can provide unicity to a subject and in six out of seven (126, or 84.6%) of sets this information was present but only in a third (47, or 37.3%) of those that information was actually used. Similarly, information on other places, schedule and social network was available in 137 (91.9%), 147 (98.7%) and 145 (97.3%) of datasets respectively, out of which were only used in 61 (44.5%), 61 (41.5%) and 37 (25.5%) respectively. These numbers highlight the availability of information past mobility research had access to which was not needed in some way, and therefore could have been left out during the data collection process. These extra cautionary steps taken in future research may help with public opinion and trust.

Given all the extra data available to researchers, any form of pre-processing beyond the change in identifiers is our next observation. For that, only one in 15 (10, or 6.7%) of sets went through some kind of pre-processing done, out of which three out of five (6, or 60.0%) had their methods clearly explained and nine in ten (9, or 90.0%) clearly stated *who* performed those extra steps. Once again, these extra bits of information help not only other researchers better understand what was done, but more importantly help assure public trust on the true aims and practices of mobility research.

Another important aspect of this analysis is how papers dealt with ethics. Even though ERB/IRB are not mandatory in most of the outlets we studied, one in six (24, or 16.1%) of papers had an IRB statement. Out of these studies, only one ([102]) provided a protocol which allows for an independent verification of that approval.

Informed consent. Given the importance of informed consent previously discussed (§ 6), we report that one in eight (19, or 12.8%) of the listed papers declared having received authorization from subjects to perform the study being presented. This goes with how people expect their data being used in accordance with the terms and services they sign before using a service, that is, an OSN user will agree in having their data being stored and processed to be used by the platform providing them the services, not some third party trying to understand changes in mobility. In most cases when consent was given, papers describe consent being given for a particular study, *e.g.*, "*All users of Locaccino, regardless of how they were recruited, gave informed consent to participate in the study work*" [44]. There are, however, exceptional cases in which a double-consent was requested to cover not only the data collection but also the specifics of the study, *e.g.*, "*Therefore, YJ performed a double consent process, where the users who have given their consent to the usage of location information and web search queries were asked again, if they wish to provide their consent to be included in the [COVID19] dataset.*" [177].

8 OUTLOOK ON HUMAN MOBILITY RESEARCH USING INDIVIDUAL DATA

In this paper, we surveyed 118 papers dating back over 20 years with the goal of reviewing how individual mobility data were used, and how methods and results were communicated with special attention to privacy and ethics. We now conclude with a series of final remarks, pushing for a future with more accountability, protecting the trust of subjects, researchers and the general public to avoid a *tragedy of the data commons* [178]. Given all the data and methods reviewed in this paper, we summarize a series of ideas and guidelines networking research using personal data could follow in the upcoming years. We do not focus on the exact steps to achieve those goals, but we leave those to the community to decide.

IMC ethical considerations. Ethical concerns have been part of network measurement community (especially IMC) for long [52]. Ethical statements are required by several scientific outlets, such as those under SIGCOMM [72], typically in the form of a required section where ethical considerations are discussed [113], and more recently, requiring that the IRB review application form be submitted along with the manuscript. Furthermore, the inclusion of ethics course as an integral part of Computer Science curricula have also been extensively discussed over the years [63, 127], fostering a growing ethical ethos among scientists. Therefore, we believe the mindset of ethically doing research with personal data should be present from the initial steps of any investigation endeavour, similar to the moral philosophy discussed by Bietti [28] for the tech industry. However, as we have highlighted with our review, a significant number of papers do not include ethical discussions even amongst recent work, let alone consider IRB reviews which need to happen *a priori* (*i.e.*, need to be considered even before a submission, not *only once this is accepted*). We also find in our review that the definitions of "ethics" and "personal data" have evolved over time. Methods to ask for consent have become more minimalist (and often implicit) so as to not inhibit a user's "quality-of-experience". Simultaneously, we observe that the ethical boundaries of large-scale (or in-the-wild) measurements are only known implicitly to researchers already well-versed

with the field. Lets take, for example, research in conducting Internet-wide scans. While there is plethora of work easily accessible via quick search on tools and methods to perform IP scanning [68]; reports on best scanning practices [53], ownership within the context of Internet [52] and guidelines for designing ethical scanning methodologies [56] are much less popularly known and sought after. Additionally, we believe a minimalist and careful handling of personal data must be practiced at all steps of the scientific process. This cautionary approach, combined with higher clarity in describing how data were used, and higher accountability are of paramount relevance for a clearer implementation of ethics in research. That is, a call for the wider adoption of the intrinsic values of ethics over the instrumental ones [28], where in the former ethics is seen as a commitment to a process while in the latter ethics is simply a means to an end (*e.g.*, fulfill requirements in a document being written). IRB reviews should be followed, and if they fail, then a redesign of the study is needed. Conferences should require a copy of the IRB check, as it is already done at IMC and other outlets.

Scientific Data Sharing Amongst the issues discussed, we believe this is one the hardest to meaningfully address with assertive measures. While on the one hand data sharing is of significant importance to the validation and reproducibility of research [47], on the other hand protecting the privacy of subjects prevents sets being shared in their original form. Current solutions require at least some level of differential privacy (see 5.2) to be met, which results in loss of information, or sharing some kind of synthetic dataset, which may still hold valuable information about the original subjects or render significantly different results from those present in the original work. Restricting research to use only shareable data are likely not the solution either. That is, limiting papers from being published only with data that have been fully anonymized and which could be shared, even under any strong NDA, cannot be a solution as it may hinder scientific development. To the problem of sharing data while protecting subject's privacy, no single rule shall be applied and the work of ethical review boards, once more, becomes even more relevant.

Ethical statements A comprehensive and thorough ethical statement must be present in all research done using individual data. While in some venues this is the norm, there are still outlets which do not enforce this as a requirement. Note, however, that is not an invitation for the inception of "ethics wash" [28] (*i.e.,* whitewash) in our community, where statements will be written only to fulfill a given requirement using *boilerplate* sentences, hiding the fact that no particular attention was paid to ethics or privacy. That is, ethical statements should not be used to cover up for bad behavior or any form of wrongdoing.

The explicit consent dilemma As we have shown, publication are using larger pools of subjects and using a growing variety of data sources. Therefore, it is naïve to expect studies with hundreds of millions of subjects will get explicit consent from every single subject, when for example, not all questions are known *a priori*, as previously discussed in the Menlo Report [18]. While no alternatives for better data governance are put in place it is the researcher's responsibility to take all cautionary steps to ensure the privacy of subjects is preserved, that all the steps taken to process and analyze those data are clearly described, and that the benefits of their research are clearly communicated for all audiences.

Ethics check committee at conferences and journals Similar to the already existing reproducibility checks, conferences and journals could implement an ethics check committee, responsible for double-checking whether essential ethics guidelines were followed (*e.g.*, ethical statement, IRB approval, informed consent). This would be available for any paper containing personal data, and as its reproducibility counterpart, would not affect the acceptance likelihood of the work. We hypothesize this *opt-in* solution could create incentives for authors to behave more ethically, and make that clear on their manuscripts, without stringent measures that could hinder scientific development in the short term.

A model for better individual mobility data governance In order to get around some of the issues discussed in this paper, we believe better traceability and accountability are plausible alternatives, but we leave the exact details of implementation for the community to decide. Regardless of the architecture of such solution, this system would be responsible for collecting, filtering, aggregating and sharing these data with internet services interested in using that information. These data points could include readings from sensors (accelerometers, GPS), contextual physical information such as proximity to others through Bluetooth, as well as online behavior. Users would then have the choice of whether or not any of those data are collected, how and if they should be aggregated or filtered, before being shared. Researchers could, in turn, publish calls for data, explaining in details their objectives and data are needed. This system could also work as a bookkeeper to all accesses to a subject's data, as well as whenever the original data are finally deleted, ensuring better accountability and possibly wider availability of data for research studies.

ACKNOWLEDGMENTS

We thank you all.

REFERENCES

[1] 2010. Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. *Pervasive and*

Mobile Computing 6, 4 (2010), 455–466. https://doi.org/10.1016/j.pmcj.2010.07.002 Human Behavior in Ubiquitous Environments: Modeling of Human Mobility Patterns.

[2] 2015. *NS Ethics '15: Proceedings of the 2015 ACM SIGCOMM Workshop on Ethics in Networked Systems Research* (London, United Kingdom). Association for Computing Machinery, New York, NY, USA.

[3] 2019. Five decades of the ACM Special Interest Group on Data Communications (SIGCOMM): A bibliometric perspective. *Computer Communication Review* 49, 5 (2019), 29–37. https://doi.org/10.1145/3371934.3371948

[4] Osman Abul, Francesco Bonchi, and Mirco Nanni. 2008. Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases. In *2008 IEEE 24th International Conference on Data Engineering*, Vol. 00. IEEE, 376–385. https://doi.org/10.1109/ICDE.2008.4497446

[5] British Academy and the Royal Society. 2017. Data management and use: governance in the 21st century.

[6] Mikhail Afanasyev, Tsuwei Chen, Geoffrey M. Voelker, and Alex C. Snoeren. 2008. Analysis of a Mixed-Use Urban Wifi Network: When Metropolitan Becomes Neapolitan. In *Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement* (Vouliagmeni, Greece) *(IMC '08)*. Association for Computing Machinery, New York, NY, USA, 85–98. https://doi.org/10.1145/1452520.1452531

[7] Laura Alessandretti, Ulf Aslak, and Sune Lehmann. 2020. The scales of human mobility. *Nature* 587, 7834 (2020), 402–407.

[8] Laura Alessandretti, Piotr Sapiezynski, Sune Lehmann, and Andrea Baronchelli. 2017. Multi-scale spatio-temporal analysis of human mobility. *PLOS ONE* 12, 2 (02 2017), 1–17. https://doi.org/10.1371/journal.pone.0171686

[9] Laura Alessandretti, Piotr Sapiezynski, Vedran Sekara, Sune Lehmann, and Andrea Baronchelli. 2018. Evidence for a conserved quantity in human mobility. *Nature human behaviour* 2, 7 (2018), 485–491.

[10] Fahad Alhasoun, May Alhazzani, Faisal Aleissa, Riyadh Alnasser, and Marta González. 2017. City scale next place prediction from sparse data through similar strangers. In *Proceedings of ACM KDD Workshop*. 191–196.

[11] Babak Alipour, Leonardo Tonetto, Aaron Yi Ding, Roozbeh Ketabi, Jörg Ott, and Ahmed Helmy. 2018. Flutes vs. Cellos: Analyzing Mobility-Traffic Correlations in Large WLAN Traces. In *2018 IEEE Conference on Computer Communications, INFOCOM 2018, Honolulu, HI, USA, April 16-19, 2018*. IEEE, 1637–1645. https://doi.org/10.1109/INFOCOM.2018.8486360

[12] Miltiadis Allamanis, Salvatore Scellato, and Cecilia Mascolo. 2012. Evolution of a location-based online social network: analysis and models. In *Proceedings of the 12th ACM SIGCOMM Internet Measurement Conference, IMC '12, Boston, MA, USA, November 14-16, 2012*, John W. Byers, Jim Kurose, Ratul Mahajan, and Alex C. Snoeren (Eds.). ACM, 145–158. https://doi.org/10.1145/2398776.2398793

[13] Mark Allman and Vern Paxson. 2007. Issues and etiquette concerning use of shared measurement data. In *Proceedings of the 7th ACM SIGCOMM Internet Measurement Conference, IMC 2007, San Diego, California, USA, October 24-26, 2007*, Constantine Dovrolis and Matthew Roughan (Eds.). ACM, 135–140. https://doi.org/10.1145/1298306.1298327

[14] Fatma Alrayes and Alia Abdelmoty. 2016. Towards Location Privacy Awareness on Geo-Social Networks. In *2016 10th International Conference on Next Generation Mobile Applications, Security and Technologies (NGMAST)*. 105–114. https://doi.org/10.1109/NGMAST.2016.26

[15] Ionut Andone, Konrad Blaszkiewicz, Matthias Böhmer, and Alexander Markowetz. 2017. Impact of location-based games on phone usage and movement: a case study on Pokémon GO. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services, MobileHCI 2017, Vienna, Austria, September 4-7, 2017*, Matt Jones, Manfred Tscheligi, Yvonne Rogers, and Roderick Murray-Smith (Eds.). ACM, 102:1–102:8. https://doi.org/10.1145/3098279.3122145

[16] Apple. 2022. About privacy and Location Services in iOS and iPadOS. Retrieved May 18, 2022 from https://support.apple.com/en-gb/HT203033

[17] James P. Bagrow, Dashun Wang, and Albert-László Barabási. 2011. Collective response of human populations to large-scale emergencies. *CoRR* abs/1106.0560 (2011). arXiv:1106.0560 http://arxiv.org/abs/1106.0560

[18] Michael Bailey, David Dittrich, Erin Kenneally, and Douglas Maughan. 2012. The Menlo Report. *IEEE Secur. Priv.* 10, 2 (2012), 71–75. https://doi.org/10.1109/MSP.2012.52

[19] Vian Bakir, Jonathan Cable, Lina Dencik, Arne Hintz, and Andrew McStay. 2015. Public feeling on privacy, security and surveillance. (2015).

[20] Magdalena Balazinska and Paul C. Castro. 2003. Characterizing Mobility and Network Usage in a Corporate Wireless Local-Area Network. In *Proceedings of the First International Conference on Mobile Systems, Applications, and Services, MobiSys 2003, San Francisco, CA, USA, May 5-8, 2003*, Daniel P. Siewiorek, Mary Baker, and Robert T. Morris (Eds.). USENIX, 303–316. https://doi.org/10.1145/1066116.1066127

[21] Michael Barbaro and Tom Zeller Jr. 2006. A Face Is Exposed for AOL Searcher No. 4417749. Retrieved May 18, 2022 from https://www.nytimes.com/2006/08/09/technology/09aol.html

[22] Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R. James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J. Ramasco, Filippo Simini, and Marcello Tomasini. 2018. Human mobility: Models and applications. *Physics Reports* 734 (2018), 1–74. https://doi.org/10.1016/j.physrep.2018.01.001 Human mobility: Models and applications.

[23] Hugo Barbosa, Fernando Buarque de Lima Neto, Alexandre G. Evsukoff, and Ronaldo Menezes. 2015. The effect of recency to human mobility. *EPJ Data Sci.* 4, 1 (2015), 21. https://doi.org/10.1140/epjds/s13688-015-0059-8

[24] BBC News. 2017. Yahoo 2013 data breach hit 'all three billion accounts'. Retrieved May 18, 2022 from https://www.bbc.com/news/business-41493494

[25] Richard A. Becker, Ramón Cáceres, Karrie J. Hanson, Sibren Isaacman, Ji Meng Loh, Margaret Martonosi, James Rowland, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. 2013. Human mobility characterization from cellular network data. *Commun. ACM* 56, 1 (2013), 74–82. https://doi.org/10.1145/2398356.2398375

[26] Richard A. Becker, Ramón Cáceres, Karrie J. Hanson, Ji Meng Loh, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. 2011. A Tale of One City: Using Cellular Network Data for Urban Planning. *IEEE Pervasive Comput.* 10, 4 (2011), 18–26. https://doi.org/10.1109/MPRV.2011.44

[27] Sourav Bhattacharya, Santi Phithakkitnukoon, Petteri Nurmi, Arto Klami, Marco Veloso, and Carlos Bento. 2013. Gaussian process-based predictive modeling for bus ridership. In *The 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '13, Zurich, Switzerland, September 8-12, 2013 - Adjunct Publication*, Friedemann Mattern, Silvia Santini, John F. Canny, Marc Langheinrich, and Jun Rekimoto (Eds.). ACM, 1189–1198. https://doi.org/10.1145/2494091.2497349

[28] Elettra Bietti. 2020. From ethics washing to ethics bashing: A view on tech ethics from within moral philosophy. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 210–219.

[29] Vincent D. Blondel, Adeline Decuyper, and Gautier Krings. 2015. A survey of results on mobile phone datasets analysis. *EPJ Data Science* 4, 1 (2015), 1–55. https://doi.org/10.1140/epjds/s13688-015-0046-0 arXiv:1502.03406

[30] Joshua E Blumenstock. 2012. Inferring patterns of internal migration from mobile phone call records: evidence from Rwanda. *Information Technology for Development* 18, 2 (2012), 107–125.

[31] Juan-José Boté and Miquel Termens. 2019. Reusing Data Technical and Ethical Challenges. *DESIDOC Journal of Library and Information Technology* 39, 06 (dec 2019), 329–337. https://doi.org/10.14429/djlit.39.06.14807

[32] Patricia Callejo, Marco Gramaglia, Rubén Cuevas, and Ángel Cuevas. 2022. A deep dive into the accuracy of IP Geolocation Databases and its impact on online advertising. *IEEE Transactions on Mobile Computing* (2022), 1–1. https://doi.org/10.1109/TMC.2022.3166785

[33] Jin Cao, Sining Chen, W. Sean Kennedy, Nicolas Kim, and Lisa Zhang. 2017. Extracting mobile user behavioral similarity via cell-level location trace. In *2017 IEEE Conference on Computer Communications Workshops, INFOCOM Workshops, Atlanta, GA, USA, May 1-4, 2017*. IEEE, 378–383. https://doi.org/10.1109/INFCOMW.2017.8116406

[34] Paul Y. Cao, Gang Li, Adam C. Champion, Dong Xuan, Steve Romig, and Wei Zhao. 2017. On human mobility predictability via WLAN logs. In *2017 IEEE Conference on Computer Communications, INFOCOM 2017, Atlanta, GA, USA, May 1-4, 2017*. IEEE, 1–9. https://doi.org/10.1109/INFOCOM.2017.8057234

[35] Augustin Chaintreau, Pan Hui, Jon Crowcroft, Christophe Diot, Richard Gass, and James Scott. 2007. Impact of Human Mobility on Opportunistic Forwarding Algorithms. *IEEE Trans. Mob. Comput.* 6, 6 (2007), 606–620. https://doi.org/10.1109/TMC.2007.1060

[36] Guangshuo Chen, Sahar Hoteit, Aline Carneiro Viana, Marco Fiore, and Carlos Sarraute. 2018. Enriching sparse mobility information in Call Detail Records. *Comput. Commun.* 122 (2018), 44–58. https://doi.org/10.1016/j.comcom.2018.03.012

[37] Rui Chen, Benjamin CM Fung, Bipin C Desai, and Nériah M Sossou. 2012. Differentially private transit data publication: a case study on the montreal transportation system. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 213–221.

[38] Zhiyuan Cheng, James Caverlee, Kyumin Lee, and Daniel Z. Sui. 2011. Exploring Millions of Footprints in Location Sharing Services. In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*, Lada A. Adamic, Ricardo Baeza-Yates, and Scott Counts (Eds.). The AAAI Press. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2783

[39] Francisco Chinchilla, Mark Lindsey, and Maria Papadopouli. 2004. Analysis of wireless information locality and association patterns in a campus. In *Proceedings IEEE INFOCOM 2004, The 23rd Annual Joint Conference of the IEEE Computer and Communications Societies, Hong Kong, China, March 7-11, 2004*. IEEE, 906–917. https://doi.org/10.1109/INFCOM.2004.1356978

[40] Eunjoon Cho, Seth A. Myers, and Jure Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011*, Chid Apté, Joydeep Ghosh, and Padhraic Smyth (Eds.). ACM, 1082–1090. https://doi.org/10.1145/2020408.2020579

[41] Seungeun Chung, Inyoung Hwang, Jiyoun Lim, and Hyun Tae Jeong. 2019. Finding Points-of-Interest (PoIs) from Life-logging and Location Trace Data. In *2019 International Conference on Information and Communication Technology Convergence (ICTC)*. 1300–1303. https://doi.org/10.1109/ICTC46691.2019.8940021

[42] David J. Crandall, Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel P. Huttenlocher, and Jon M. Kleinberg. 2010. Inferring social ties from geographic coincidences. *Proc. Natl. Acad. Sci. USA* 107, 52 (2010), 22436–22441. https://doi.org/10.1073/pnas.1006155107

[43] David J. Crandall, Lars Backstrom, Daniel P. Huttenlocher, and Jon M. Kleinberg. 2009. Mapping the world's photos. In *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*, Juan Quemada, Gonzalo León, Yoëlle S. Maarek, and Wolfgang Nejdl (Eds.). ACM, 761–770. https://doi.org/10.1145/1526709.1526812

[44] Justin Cranshaw, Eran Toch, Jason I. Hong, Aniket Kittur, and Norman M. Sadeh. 2010. Bridging the gap between physical location and online social networks. In *UbiComp 2010: Ubiquitous Computing, 12th International Conference, UbiComp 2010, Copenhagen, Denmark, September 26-29, 2010, Proceedings (ACM International Conference Proceeding Series)*, Jakob E. Bardram, Marc Langheinrich, Khai N. Truong, and Paddy Nixon (Eds.). ACM, 119–128. https://doi.org/10.1145/1864349.1864380

[45] Balázs Csanád Csáji, Arnaud Browet, Vincent A. Traag, Jean-Charles Delvenne, Etienne Huens, Paul Van Dooren, Zbigniew Smoreda, and Vincent D. Blondel. 2012. Exploring the Mobility of Mobile Phone Users. *CoRR* abs/1211.6014 (2012). arXiv:1211.6014 http://arxiv.org/abs/1211.6014

[46] Yves Alexandre De Montjoye, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. 2013. Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports* 3 (2013), 1–5. https://doi.org/10.1038/srep01376

[47] Nurullah Demir, Matteo Große-Kampmann, Tobias Urban, Christian Wressnegger, Thorsten Holz, and Norbert Pohlmann. 2022. Reproducibility and Replicability of Web Measurement Studies. In *Proceedings of the ACM Web Conference 2022*, Vol. 1. ACM, New York, NY, USA, 533–544. https://doi.org/10.1145/3485447.3512214

[48] Education Department of Health and Welfare. 1979. *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research.* https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html

[49] Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R Stevens, Andrea E Gaughan, Vincent D Blondel, and Andrew J Tatem. 2014. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences* 111, 45 (2014), 15888–15893.

[50] Pierre Deville, Chaoming Song, Nathan Eagle, Vincent D. Blondel, Albert-László Barabási, and Dashun Wang. 2016. Scaling identity connects human mobility and social interactions. *Proceedings of the National Academy of Sciences* 113, 26 (2016), 7047–7052. https://doi.org/10.1073/pnas.1525443113 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1525443113

[51] Riccardo Di Clemente, Miguel Luengo-Oroz, Matias Travizano, Sharon Xu, Bapu Vaitla, and Marta C González. 2018. Sequences of purchases in credit card data reveal lifestyles in urban populations. *Nature communications* 9, 1 (2018), 1–8.

[52] Sven Dietrich, Jeroen van der Ham, Aiko Pras, Roland van Rijswijk-Deij, Darren Shou, Anna Sperotto, Aimee van Wynsberghe, and Lenore D. Zuck. 2014. Ethics in Data Sharing: Developing a Model for Best Practice. In *35. IEEE Security and Privacy Workshops, SPW 2014, San Jose, CA, USA, May 17-18, 2014*. IEEE Computer Society, 5–9. https://doi.org/10.1109/SPW.2014.43

[53] David Dittrich et al. 2012. The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research. *US DHS* (2012).

[54] Trinh Minh Tri Do and Daniel Gatica-Perez. 2012. Contextual conditional models for smartphone-based human mobility prediction. In *The 2012 ACM Conference on Ubiquitous Computing, Ubicomp '12, Pittsburgh, PA, USA, September 5-8, 2012*, Anind K. Dey, Hao-Hua Chu, and Gillian R. Hayes (Eds.). ACM, 163–172. https://doi.org/10.1145/2370216.2370242

[55] Matt Duckham and Lars Kulik. 2005. A formal model of obfuscation and negotiation for location privacy. In *International conference on pervasive computing*. Springer, 152–170.

[56] Zakir Durumeric, Eric Wustrow, and J. Alex Halderman. 2013. ZMap: Fast Internet-Wide Scanning and Its Security Applications. In *Proceedings of the 22nd USENIX Conference on Security* (Washington, D.C.) *(SEC'13)*. USENIX Association, USA, 605–620.

[57] Cynthia Dwork. 2008. Differential Privacy: A Survey of Results. In *Theory and Applications of Models of Computation*. Vol. 4978 LNCS. Springer Berlin Heidelberg, Berlin, Heidelberg, 1–19. https://doi.org/10.1007/978-3-540-79228-4_1

[58] Nathan Eagle, Aaron Clauset, and John A. Quinn. 2009. Location Segmentation, Inference and Prediction for Anticipatory Computing. In *Technosocial Predictive Analytics, Papers from the 2009 AAAI Spring Symposium, Technical Report SS-09-09, Stanford, California, USA, March 23-25, 2009*. AAAI, 20–25. http://www.aaai.org/Library/Symposia/Spring/2009/ss09-09-005.php

[59] Nathan Eagle and Alex Pentland. 2006. Reality mining: sensing complex social systems. *Pers. Ubiquitous Comput.* 10, 4 (2006), 255–268. https://doi.org/10.1007/s00779-005-0046-3

[60] Zhihan Fang, Fan Zhang, Ling Yin, and Desheng Zhang. 2018. MultiCell: Urban Population Modeling Based on Multiple Cellphone Networks. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3 (2018), 106:1–106:25. https://doi.org/10.1145/3264916

[61] Jie Feng, Yong Li, Chao Zhang, Funing Sun, Fanchao Meng, Ang Guo, and Depeng Jin. 2018. DeepMove: Predicting Human Mobility with Attentional Recurrent Networks. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, Pierre-Antoine Champin, Fabien Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis (Eds.). ACM, 1459–1468. https://doi.org/10.1145/3178876.3186058

[62] Marco Fiore, Panagiota Katsikouli, Elli Zavou, Mathieu Cunche, Françoise Fessant, Dominique Le Hello, Ulrich Aivodji, Baptiste Olivier, Tony Quertier, and Razvan Stanica. 2020. Privacy in trajectory micro-data publishing: a survey. *Transactions on Data Privacy* 13 (2020), 91–149.

[63] Task Force. 2020. Computing Curricula 2020: Paradigms for global computing education. *Assoc. Comput. Mach.(ACM) IEEE Comput. Comput.(IEEE-CS), New York, NY, USA, Rep* 10 (2020), 3467967.

[64] aline shakti franzke, Anja Bechmann, Michael Zimmer, and Charles Ess. 2020. Internet Research: Ethical Guidelines 3.0. *Association of Internet Researchers* (2020).

[65] Julien Freudiger, Reza Shokri, and Jean-Pierre Hubaux. 2012. Evaluating the Privacy Risk of Location-Based Services. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 7035 LNCS. 31–46. https://doi.org/10.1007/978-3-642-27576-0_3

[66] Benjamin CM Fung, Ke Wang, Rui Chen, and Philip S Yu. 2010. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (Csur)* 42, 4 (2010), 1–53.

[67] Laetitia Gauvin, Michele Tizzoni, Simone Piaggesi, Andrew Young, Natalia Adler, Stefaan Verhulst, Leo Ferres, and Ciro Cattuto. 2020. Gender gaps in urban mobility. *Humanities and Social Sciences Communications* 7, 1 (2020), 1–13.

[68] GeekFlare. 2022. *11 Best IP Scanner Tools for Network Management*. https://geekflare.com/network-scanner/

[69] John R. Giles, Elisabeth zu Erbach-Schoenberg, Andrew J. Tatem, Lauren Gardner, Ottar N. Bjørnstad, C. J. E. Metcalf, and Amy Wesolowski. 2020. The duration of travel impacts the spatial dynamics of infectious diseases. *Proceedings of the National Academy of Sciences* 117, 36 (2020), 22572–22579. https://doi.org/10.1073/pnas.1922663117 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1922663117

[70] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. 2008. Understanding individual human mobility patterns. *nature* 453, 7196 (2008), 779–782. https://doi.org/10.1038/nature06958

[71] Google. 2022. Manage your Android device's location settings. Retrieved May 18, 2022 from https://support.google.com/accounts/answer/3467281?hl=en

[72] DW Gotterbarn, Bo Brinkman, Catherine Flick, Michael S Kirkpatrick, Keith Miller, Kate Vazansky, and Marty J Wolf. 2018. ACM code of ethics and professional conduct. (2018).

[73] Matthieu Gouel, Kevin Vermeulen, Robert Beverly, Olivier Fourmaux, and Timur Friedman. 2021. IP Geolocation Database Stability and Implications for Network Research: A Reproducibility Study. In *5th Network Traffic Measurement and Analysis Conference, TMA 2021, Virtual Event, September 14-15, 2021*, Vaibhav Bajpai, Hamed Haddadi, and Oliver Hohlfeld (Eds.). IFIP. http://dl.ifip.org/db/conf/tma/tma2021/tma2021-paper2.pdf

[74] Mehmet Emre Gursoy, Ling Liu, Stacey Truex, and Lei Yu. 2018. Differentially private and utility preserving publication of trajectory data. *IEEE Transactions on Mobile Computing* 18, 10 (2018), 2315–2329.

[75] Georg Heiler, Tobias Reisch, Jan Hurt, Mohammad Forghani, Aida Omani, Allan Hanbury, and Farid Karimipour. 2020. Country-wide Mobility Changes Observed Using Mobile Phone Data During COVID-19 Pandemic. In *2020 IEEE International Conference on Big Data (IEEE BigData 2020), Atlanta, GA, USA, December 10-13, 2020*, Xintao Wu, Chris Jermaine, Li Xiong, Xiaohua Hu, Olivera Kotevska, Siyuan Lu, Weija Xu, Srinivas Aluru, Chengxiang Zhai, Eyhab Al-Masri, Zhiyuan Chen, and Jeff Saltz (Eds.). IEEE, 3123–3132. https://doi.org/10.1109/BigData50022.2020.9378374

[76] Tristan Henderson, David Kotz, and Ilya Abyzov. 2004. The changing usage of a mature campus-wide wireless network. In *Proceedings of the 10th Annual International Conference on Mobile Computing and Networking, MOBICOM 2004, 2004, Philadelphia, PA, USA, September 26 - October 1, 2004*, Zygmunt J. Haas, Samir R. Das, and Ravi Jain (Eds.). ACM, 187–201. https://doi.org/10.1145/1023720.1023739

[77] Desislava Hristova, Matthew J. Williams, Mirco Musolesi, Pietro Panzarasa, and Cecilia Mascolo. 2016. Measuring Urban Social Diversity Using Interconnected Geo-Social Networks. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, Jacqueline Bourdeau, Jim Hendler, Roger Nkambou, Ian Horrocks, and Ben Y. Zhao (Eds.). ACM, 21–30. https://doi.org/10.1145/2872427.2883065

[78] Jie Huang, David Levinson, Jiaoe Wang, Jiangping Zhou, and Zi jia Wang. 2018. Tracking job and housing dynamics with smartcard data. *Proceedings of the National Academy of Sciences* 115, 50 (2018), 12710–12715. https://doi.org/10.1073/pnas.1815928115 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1815928115

[79] Xiao Huang, Zhenlong Li, Yuqin Jiang, Xinyue Ye, Chengbin Deng, Jiajia Zhang, and Xiaoming Li. 2021. The characteristics of multi-source mobility datasets and how they reveal the luxury nature of social distancing in the US during the COVID-19 pandemic. *International Journal of Digital Earth* 14, 4 (2021), 424–442.

[80] Pan Hui, Augustin Chaintreau, James Scott, Richard Gass, Jon Crowcroft, and Christophe Diot. 2005. Pocket switched networks and human mobility in conference environments. In *Proceedings*

of the 2005 ACM SIGCOMM Workshop on Delay-Tolerant Networking, WDTN '05, Philadelphia, Pennsylvania, USA, August 26, 2005, Kevin Fall and Srinivasan Keshav (Eds.). ACM, 244–251. https://doi.org/10.1145/1080139.1080142

[81] Ron Hutchins and Ellen W. Zegura. 2002. Measurements from a campus wireless network. In *IEEE International Conference on Communications, ICC 2002, April 28 - May 2, 2002, New York City, NY, USA*. IEEE, 3161–3167. https://doi.org/10.1109/ICC.2002.997419

[82] Marcello Ienca and Effy Vayena. 2021. Ethical requirements for responsible research with hacked data. *Nature Machine Intelligence* 3, 9 (2021), 744–748.

[83] Sibren Isaacman, Richard A. Becker, Ramón Cáceres, Stephen G. Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. 2011. Ranges of human mobility in Los Angeles and New York. In *Ninth Annual IEEE International Conference on Pervasive Computing and Communications, PerCom 2011, 21-25 March 2011, Seattle, WA, USA, Workshop Proceedings*. IEEE Computer Society, 88–93. https://doi.org/10.1109/PERCOMW.2011.5766977

[84] Olle Järv, Ago Tominga, Kerli Müürisepp, and Siiri Silm. 2021. The impact of COVID-19 on daily lives of transnational people based on smartphone data: Estonians in Finland. *J. Locat. Based Serv.* 15, 3 (2021), 169–197. https://doi.org/10.1080/17489725.2021.1887526

[85] Bjørn Sand Jensen, Jakob Eg Larsen, Kristian Jensen, Jan Larsen, and Lars Kai Hansen. 2010. Estimating human predictability from mobile sensor data. In *2010 IEEE International Workshop on Machine Learning for Signal Processing*. 196–201. https://doi.org/10.1109/MLSP.2010.5588997

[86] Shan Jiang, Joseph Ferreira Jr., and Marta C. González. 2017. Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore. *IEEE Trans. Big Data* 3, 2 (2017), 208–219. https://doi.org/10.1109/TBDATA.2016.2631141

[87] Shan Jiang, Yingxiang Yang, Siddharth Gupta, Daniele Veneziano, Shounak Athavale, and Marta C. González. 2016. The TimeGeo modeling framework for urban mobility without travel surveys. *Proceedings of the National Academy of Sciences* 113, 37 (2016), E5370–E5378. https://doi.org/10.1073/pnas.1524261113 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1524261113

[88] Raja Jurdak, Kun Zhao, Jiajun Liu, Maurice AbouJaoude, Mark Cameron, and David Newth. 2015. Understanding human mobility from Twitter. *PLoS ONE* 10, 7 (2015), 35. https://doi.org/10.1371/journal.pone.0131469 arXiv:arXiv:1412.2154v2

[89] Chaogui Kang, Xiujun Ma, Daoqin Tong, and Yu Liu. 2012. Intra-urban human mobility patterns: An urban morphology perspective. *Physica A: Statistical Mechanics and its Applications* 391, 4 (2012), 1702–1717. https://doi.org/10.1016/j.physa.2011.11.005

[90] David Kotz and Kobby Essien. 2005. Analysis of a Campus-Wide Wireless Network. *Wirel. Networks* 11, 1-2 (2005), 115–133. https://doi.org/10.1007/s11276-004-4750-0

[91] David Kotz and Tristan Henderson. 2005. Crawdad: A community resource for archiving wireless data at dartmouth. *IEEE Pervasive Computing* 4, 4 (2005), 12–14.

[92] Moritz UG Kraemer, Adam Sadilek, Qian Zhang, Nahema A Marchal, Gaurav Tuli, Emily L Cohn, Yulin Hswen, T Alex Perkins, David L Smith, Robert C Reiner, et al. 2020. Mapping global variation in human mobility. *Nature Human Behaviour* 4, 8 (2020), 800–810.

[93] Stuart Kurkowski, Tracy Camp, and Michael Colagrosso. 2005. MANET simulation studies: The Incredibles. *ACM SIGMOBILE Mobile Computing and Communications Review* 9, 4 (oct 2005), 50–61. https://doi.org/10.1145/1096166.1096174

[94] Kevin Lai, Mema Roussopoulos, Diane Tang, Xinhua Zhao, and Mary Baker. 1998. Experiences with a Mobile Testbed. In *Worldwide Computing and Its Applications, International Conference, WWCA '98, Second International Conference, Tsukuba, Japan, March 4-5, 1998, Proceedings (Lecture Notes in Computer Science, Vol. 1368)*, Yoshifumi Masunaga, Takuya Katayama, and Michiharu Tsukamoto (Eds.). Springer, 222–237. https://doi.org/10.1007/3-540-64216-1_51

[95] Neal Lathia and Licia Capra. 2011. How smart is your smartcard?: measuring travel behaviours, perceptions, and incentives. In *UbiComp 2011: Ubiquitous Computing, 13th International Conference, UbiComp 2011, Beijing, China, September 17-21, 2011, Proceedings*, James A. Landay, Yuanchun Shi, Donald J. Patterson, Yvonne Rogers, and Xing Xie (Eds.). ACM, 291–300. https://doi.org/10.1145/2030112.2030152

[96] Neal Lathia, Daniele Quercia, and Jon Crowcroft. 2012. The Hidden Image of the City: Sensing Community Well-Being from Urban Mobility. In *Pervasive Computing - 10th International Conference, Pervasive 2012, Newcastle, UK, June 18-22, 2012. Proceedings (Lecture Notes in Computer Science, Vol. 7319)*, Judy Kay, Paul Lukowicz, Hideyuki Tokuda, Patrick Olivier, and Antonio Krüger (Eds.). Springer, 91–98. https://doi.org/10.1007/978-3-642-31205-2_6

[97] Maxime Lenormand, Bruno Gonçalves, Antonia Tugores, and José J Ramasco. 2015. Human diffusion and city influence. *Journal of The Royal Society Interface* 12, 109 (2015), 20150473.

[98] Shan Lu, Jichang Zhao, and Huiwen Wang. 2022. Academic failures and co-location social networks in campus. *EPJ Data Science* 11, 1 (2022). https://doi.org/10.1140/epjds/s13688-022-00322-0

[99] Xin Lu, Linus Bengtsson, and Petter Holme. 2012. Predictability of population displacement after the 2010 Haiti earthquake. *Proceedings of the National Academy of Sciences* 109, 29 (2012), 11576–11581.

[100] Xin Lu, Erik Wetter, Nita Bharti, Andrew J Tatem, and Linus Bengtsson. 2013. Approaching the limit of predictability in human mobility. *Scientific reports* 3, 1 (2013), 1–9.

[101] Massimiliano Luca, Gianni Barlacchi, Bruno Lepri, and Luca Pappalardo. 2021. A Survey on Deep Learning for Human Mobility. *ACM Comput. Surv.* 55, 1, Article 7 (nov 2021), 44 pages. https://doi.org/10.1145/3485125

[102] Shaojun Luo, Flaviano Morone, Carlos Sarraute, Matías Travizano, and Hernán A Makse. 2017. Inferring personal economic status from social network location. *Nature Communications* 8, 1 (2017), 1–7.

[103] Andra Lutu, Diego Perino, Marcelo Bagnulo, Enrique Frías-Martínez, and Javad Khangosstar. 2020. A Characterization of the COVID-19 Pandemic Impact on a Mobile Network Operator Traffic. In *IMC '20: ACM Internet Measurement Conference, Virtual Event, USA, October 27-29, 2020*. ACM, 19–33. https://doi.org/10.1145/3419394.3423655

[104] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1, 1 (2007), 3–es.

[105] Marvin McNett and Geoffrey M. Voelker. 2005. Access and mobility of wireless PDA users. *ACM SIGMOBILE Mob. Comput. Commun. Rev.* 9, 2 (2005), 40–55. https://doi.org/10.1145/1072989.1072995

[106] Saeed Moghaddam and Ahmed Helmy. 2011. Multidimensional modeling and analysis of wireless users online activity and mobility: a neural-networks map approach. In *Proceedings of the 14th International Symposium on Modeling Analysis and Simulation of Wireless and Mobile Systems, MSWiM 2011, Miami, Florida, USA, October 31 - November 4, 2011*, Ahmed Helmy, Björn Landfeldt, and Luciano Bononi (Eds.). ACM, 401–408. https://doi.org/10.1145/2068897.2068965

[107] Anastasios Noulas, Salvatore Scellato, Renaud Lambiotte, Massimiliano Pontil, and Cecilia Mascolo. 2011. A tale of many cities: universal patterns in human urban mobility. *CoRR* abs/1108.5355 (2011). arXiv:1108.5355 http://arxiv.org/abs/1108.5355

[108] State of California Department of Justice. 2018. California Consumer Privacy Act (CCPA). Retrieved May 18, 2022 from https://oag.ca.gov/privacy/ccpa

[109] Association of Computing Machinery (ACM). 2021. *ACM Publications Policy on Research Involving Human Participants and Subjects.* https://www.acm.org/publications/policies/research-involving-human-participants-and-subjects

[110] Eduardo Mucelli Rezende Oliveira, Aline Carneiro Viana, Carlos Sarraute, Jorge Brea, and J. Ignacio Alvarez-Hamelin. 2016. On the regularity of human mobility. *Pervasive Mob. Comput.* 33 (2016), 73–90. https://doi.org/10.1016/j.pmcj.2016.04.005

[111] Luca Pappalardo, Leo Ferres, Manuel Sacasa, Ciro Cattuto, and Loreto Bravo. 2021. Evaluation of home detection algorithms on mobile phone data using individual-level ground truth. *EPJ Data Sci.* 10, 1 (2021), 29. https://doi.org/10.1140/epjds/s13688-021-00284-9

[112] Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-László Barabási. 2015. Returners and explorers dichotomy in human mobility. *Nature communications* 6, 1 (2015), 1–8.

[113] Craig Partridge and Mark Allman. 2016. Ethical considerations in network measurement papers. *Commun. ACM* 59, 10 (2016), 58–64. https://doi.org/10.1145/2896816

[114] Santi Phithakkitnukoon, Francesco Calabrese, Zbigniew Smoreda, and Carlo Ratti. 2011. Out of Sight Out of Mind-How Our Mobile Social Network Changes during Migration. In *PASSAT/SocialCom 2011, Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom), Boston, MA, USA, 9-11 Oct., 2011.* IEEE Computer Society, 515–520. https://doi.org/10.1109/PASSAT/SocialCom.2011.11

[115] Santi Phithakkitnukoon, Teerayut Horanont, Giusy Di Lorenzo, Ryosuke Shibasaki, and Carlo Ratti. 2010. Activity-Aware Map: Identifying Human Daily Activity Pattern Using Mobile Phone Data. In *Human Behavior Understanding, First International Workshop, HBU 2010, Istanbul, Turkey, August 22, 2010. Proceedings (Lecture Notes in Computer Science, Vol. 6219)*, Albert Ali Salah, Theo Gevers, Nicu Sebe, and Alessandro Vinciarelli (Eds.). Springer, 14–25. https://doi.org/10.1007/978-3-642-14715-9_3

[116] Santi Phithakkitnukoon and Carlo Ratti. 2011. Inferring asymmetry of inhabitant flow using call detail records. (2011).

[117] Nicolas B. Ponieman, Alejo Salles, and Carlos Sarraute. 2013. Human mobility and predictability enriched by social phenomena information. In *Advances in Social Networks Analysis and Mining 2013, ASONAM '13, Niagara, ON, Canada - August 25 - 29, 2013*, Jon G. Rokne and Christos Faloutsos (Eds.). ACM, 1331–1336. https://doi.org/10.1145/2492517.2500236

[118] Kangjie Lu Qiushi Wu. 2021. *Retraction – On the Feasibilityof Stealthily Introducing Vulnerabilities inOpen-Source Software via Hypocrite Commits.* https://www-users.cse.umn.edu/~kjlu/papers/withdrawal-letter.pdf

[119] Daniele Quercia, Neal Lathia, Francesco Calabrese, Giusy Di Lorenzo, and Jon Crowcroft. 2010. Recommending Social Events from Mobile Phone Location Data. In *ICDM 2010, The 10th IEEE International Conference on Data Mining, Sydney, Australia, 14-17 December 2010*, Geoffrey I. Webb, Bing Liu, Chengqi Zhang, Dimitrios Gunopulos, and Xindong Wu (Eds.). IEEE Computer Society, 971–976. https://doi.org/10.1109/ICDM.2010.152

[120] Gyan Ranjan, Hui Zang, Zhi-Li Zhang, and Jean Bolot. 2012. Are call detail records biased for sampling human mobility? *ACM SIGMOBILE Mob. Comput. Commun. Rev.* 16, 3 (2012), 33–44. https://doi.org/10.1145/2412096.2412101

[121] Injong Rhee, Minsu Shin, Seongik Hong, Kyunghan Lee, and Song Chong. 2008. On the Levy-Walk Nature of Human Mobility. In *INFOCOM 2008. 27th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies, 13-18 April 2008, Phoenix, AZ, USA.* IEEE, 924–932. https://doi.org/10.1109/INFOCOM.2008.145

[122] Luca Rossi, James Walker, and Mirco Musolesi. 2015. Spatio-temporal techniques for user identification by means of GPS mobility data. *EPJ Data Science* 4, 1 (2015), 11.

[123] Adam Sadilek, Henry Kautz, and Jeffrey P. Bigham. 2012. Finding your friends and following them to where you are. *WSDM 2012 - Proceedings of the 5th ACM International Conference on Web Search and Data Mining* (2012), 723–732. https://doi.org/10.1145/2124295.2124380

[124] Adam Sadilek and John Krumm. 2012. Far Out: Predicting Long-Term Human Mobility. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Ontario, Canada*, Jörg Hoffmann and Bart Selman (Eds.). AAAI Press. http://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/4845

[125] Amin Sadri, Flora D. Salim, Yongli Ren, Wei Shao, John Krumm, and Cecilia Mascolo. 2018. What Will You Do for the Rest of the Day?: An Approach to Continuous Trajectory Prediction. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4 (2018), 186:1–186:26. https://doi.org/10.1145/3287064

[126] Julián Salas, David Megías, Vicenç Torra, Marina Toger, Joel Dahne, and Raazesh Sainudiin. 2020. Swapping trajectories with a sufficient sanitizer. *Pattern Recognition Letters* 131 (2020), 474–480.

[127] Jeffrey S Saltz, Neil I Dewar, and Robert Heckman. 2018. Key concepts for a data science ethics curriculum. In *Proceedings of the 49th ACM technical symposium on computer science education.* 952–957.

[128] Salvatore Scellato and Cecilia Mascolo. 2011. Measuring user activity on an online location-based social network. In *2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS).* 918–923. https://doi.org/10.1109/INFCOMW.2011.5928943

[129] D. Schwab and R. Bunt. 2004. Characterising the use of a campus wireless network. In *IEEE INFOCOM 2004*, Vol. 2. 862–870 vol.2. https://doi.org/10.1109/INFCOM.2004.1356974

[130] searchids. 2006. Search-ID. Retrieved May 18, 2022 from http://searchids.com/

[131] Yohei Shida, Hideki Takayasu, Shlomo Havlin, and Misako Takayasu. 2021. Universal scaling of human flow remain unchanged during the COVID-19 pandemic. *Appl. Netw. Sci.* 6, 1 (2021), 75. https://doi.org/10.1007/s41109-021-00416-0

[132] Reza Shokri, George Theodorakopoulos, Carmela Troncoso, Jean-Pierre Hubaux, and Jean-Yves Le Boudec. 2012. Protecting location privacy. In *Proceedings of the 2012 ACM conference on Computer and communications security - CCS '12.* ACM Press, New York, New York, USA, 617. https://doi.org/10.1145/2382196.2382261

[133] Filippo Simini, Marta C González, Amos Maritan, and Albert-László Barabási. 2012. A universal model for mobility and migration patterns. *Nature* 484, 7392 (2012), 96–100.

[134] The Royal Society. 2017. Data governance: public engagement review. *British Academy* (2017).

[135] Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. 2010. Modelling the scaling properties of human mobility. *Nature physics* 6, 10 (2010), 818–823.

[136] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi. 2010. Limits of Predictability in Human Mobility. *Science* 327, 5968 (2010), 1018–1021. https://doi.org/10.1126/science.1177170 arXiv:0307014 [cond-mat]

[137] Xuan Song, Quanshi Zhang, Yoshihide Sekimoto, Teerayut Horanont, Satoshi Ueyama, and Ryosuke Shibasaki. 2013. Modeling and probabilistic reasoning of population evacuation during large-scale disaster. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*, Inderjit S. Dhillon, Yehuda Koren, Rayid Ghani, Ted E. Senator, Paul Bradley, Rajesh Parekh, Jingrui He, Robert L.

Grossman, and Ramasamy Uthurusamy (Eds.). ACM, 1231–1239. https://doi.org/10.1145/2487575.2488189

[138] Yi Song, Daniel Dahlmeier, and Stéphane Bressan. 2014. Not So Unique in the Crowd: a Simple and Effective Algorithm for Anonymizing Location Data. In *Proceeding of the 1st International Workshop on Privacy-Preserving IR: When Information Retrieval Meets Privacy and Security co-located with 37th Annual International ACM SIGIR conference, PIR@SIGIR 2014, Gold Coast, Australia, July 11, 2014 (CEUR Workshop Proceedings, Vol. 1225)*, Luo Si and Hui Yang (Eds.). CEUR-WS.org, 19–24. http://ceur-ws.org/Vol-1225/pir2014_submission_11.pdf

[139] Mudhakar Srivatsa and Mike Hicks. 2012. Deanonymizing mobility traces. In *Proceedings of the 2012 ACM conference on Computer and communications security - CCS '12*. ACM Press, New York, New York, USA, 628. https://doi.org/10.1145/2382196.2382262

[140] Arkadiusz Stopczynski, Vedran Sekara, Piotr Sapiezynski, Andrea Cuttone, Mette My Madsen, Jakob Eg Larsen, and Sune Lehmann. 2014. Measuring large-scale social networks with high resolution. *PLoS ONE* 9, 4 (2014). https://doi.org/10.1371/journal.pone.0095978 arXiv:1401.7233

[141] Lijun Sun, Kay W Axhausen, Der-Horng Lee, and Manuel Cebrian. 2014. Efficient detection of contagious outbreaks in massive metropolitan encounter networks. *Scientific reports* 4, 1 (2014), 1–6.

[142] Lijun Sun, Kay W. Axhausen, Der-Horng Lee, and Xianfeng Huang. 2013. Understanding metropolitan patterns of daily encounters. *Proceedings of the National Academy of Sciences* 110, 34 (2013), 13774–13779. https://doi.org/10.1073/pnas.1306440110 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1306440110

[143] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002), 557–570. https://doi.org/10.1142/S0218488502001648

[144] Diane Tang and Mary Baker. 1999. Analysis of a Metropolitan-Area Wireless Network. In *MOBICOM '99, The Fifth Annual ACM/IEEE International Conference on Mobile Computing and Networking, Seattle, Washington, USA, August 15-19, 1999*, Harel Kodesh, Victor Bahl, Tomasz Imielinski, and Martha Steenstrup (Eds.). ACM, 13–23. https://doi.org/10.1145/313451.313460

[145] Diane Tang and Mary Baker. 2000. Analysis of a local-area wireless network. In *MOBICOM 2000, Proceedings of the sixth annual international conference on Mobile computing and networking, Boston, MA, USA, August 6-11, 2000*, Raymond L. Pickholtz, Sajal K. Das, Ramón Cáceres, and J. J. Garcia-Luna-Aceves (Eds.). ACM, 1–10. https://doi.org/10.1145/345910.345912

[146] Jinjun Tang, Fang Liu, Yinhai Wang, and Hua Wang. 2015. Uncovering urban human mobility from large scale taxi GPS data. *Physica A: Statistical Mechanics and its Applications* 438 (2015), 140–153. https://doi.org/10.1016/j.physa.2015.06.032

[147] Andrew J Tatem, Youliang Qiu, David L Smith, Oliver Sabot, Abdullah S Ali, and Bruno Moonen. 2009. The use of mobile phone data for the estimation of the travel patterns and imported Plasmodium falciparum rates among Zanzibar residents. *Malaria journal* 8, 1 (2009), 1–12.

[148] Douglas do Couto Teixeira, Aline Carneiro Viana, Mário S. Alvim, and Jussara M. Almeida. 2019. Deciphering Predictability Limits in Human Mobility *(SIGSPATIAL '19)*. Association for Computing Machinery, New York, NY, USA, 52–61. https://doi.org/10.1145/3347146.3359093

[149] Manolis Terrovitis and Nikos Mamoulis. 2008. Privacy preservation in the publication of trajectories. In *The Ninth international conference on mobile data management (mdm 2008)*. IEEE, 65–72.

[150] Suttipong Thajchayapong and Jon M. Peha. 2003. Mobility patterns in microcellular wireless networks. In *2003 IEEE Wireless Communications and Networking, WCNC 2003, New Orleans, LA, USA, 16-20 March, 2003*. IEEE, 1963–1968. https://doi.org/10.1109/WCNC.2003.1200688

[151] Kanchana Thilakarathna, Suranga Seneviratne, Kamal Gupta, Mohamed Ali Kâafar, and Aruna Seneviratne. 2017. A deep dive into location-based communities in social discovery networks. *Comput. Commun.* 100 (2017), 78–90. https://doi.org/10.1016/j.comcom.2016.11.008

[152] Daniel R. Thomas, Sergio Pastrana, Alice Hutchings, Richard Clayton, and Alastair R. Beresford. 2017. Ethical issues in research using datasets of illicit origin. *Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC* (2017), 445–462.

[153] Etienne Thuillier, Laurent Moalic, Sid Lamrous, and Alexandre Caminada. 2018. Clustering Weekly Patterns of Human Mobility Through Mobile Phone Data. *IEEE Trans. Mob. Comput.* 17, 4 (2018), 817–830. https://doi.org/10.1109/TMC.2017.2742953

[154] The New York Times. 2020. How the Virus Got Out. Retrieved May 18, 2022 from https://www.nytimes.com/interactive/2020/03/22/world/coronavirus-spread.html

[155] Jameson L Toole, Carlos Herrera-Yaqüe, Christian M Schneider, and Marta C González. 2015. Coupling human mobility and social ties. *Journal of The Royal Society Interface* 12, 105 (2015), 20141128.

[156] Wellcome Trust. 2013. Summary report of qualitative research into public attitudes to personal data and linking personal data.

[157] Kota Tsubouchi, Tomoki Saito, and Masamichi Shimosaka. 2019. Context-Based Markov Model toward Spatio-Temporal Prediction with Realistic Dataset *(PredictGIS'19)*. Association for Computing Machinery, New York, NY, USA, 24–32. https://doi.org/10.1145/3356995.3364534

[158] Cristian Tuduce and Thomas R. Gross. 2005. A mobility model based on WLAN traces and its validation. In *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies, 13-17 March 2005, Miami, FL, USA*. IEEE, 664–674. https://doi.org/10.1109/INFCOM.2005.1497932

[159] Jeroen van der Ham. 2017. Ethics and Internet Measurements. In *2017 IEEE Security and Privacy Workshops, SP Workshops 2017, San Jose, CA, USA, May 25, 2017*. IEEE Computer Society, 247–251. https://doi.org/10.1109/SPW.2017.17

[160] Sudip Vhaduri and Christian Poellabauer. 2016. Cooperative Discovery of Personal Places from Location Traces. In *25th International Conference on Computer Communication and Networks, ICCCN 2016, Waikoloa, HI, USA, August 1-4, 2016*. IEEE, 1–9. https://doi.org/10.1109/ICCCN.2016.7568500

[161] Sudip Vhaduri, Christian Poellabauer, Aaron Striegel, Omar Lizardo, and David Hachen. 2017. Discovering places of interest using sensor data from smartphones and wearables. In *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation, SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI 2017, San Francisco, CA, USA, August 4-8, 2017*. IEEE, 1–8. https://doi.org/10.1109/UIC-ATC.2017.8397495

[162] Long H. Vu, Phuong Nguyen, Klara Nahrstedt, and Björn Richerzhagen. 2015. Characterizing and modeling people movement from mobile phone sensing traces. *Pervasive Mob. Comput.* 17 (2015), 220–235. https://doi.org/10.1016/j.pmcj.2014.12.001

[163] Dashun Wang and Chaoming Song. 2015. Impact of human mobility on social networks. *J. Commun. Networks* 17, 2 (2015), 100–109. https://doi.org/10.1109/JCN.2015.000023

[164] Huandong Wang, Chen Gao, Yong Li, Gang Wang, Depeng Jin, and Jingbo Sun. 2018. De-anonymization of Mobility Trajectories: Dissecting the Gaps between Theory and Practice. In *Proceedings 2018 Network and Distributed System Security Symposium*, Vol. 20. Internet Society, Reston, VA, 796–815. https://doi.org/10.14722/ndss.2018.23211

[165] Huandong Wang, Yong Li, Sihan Zeng, Gang Wang, Pengyu Zhang, Pan Hui, and Depeng Jin. 2019. Modeling Spatio-Temporal App Usage for a Large User Population. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 1 (2019), 27:1–27:23. https://doi.org/10.1145/3314414

[166] Qi Wang, Nolan Edward Phillips, Mario L. Small, and Robert J. Sampson. 2018. Urban mobility and neighborhood isolation in America’s 50 largest cities. *Proceedings of the National Academy of Sciences* 115, 30 (2018), 7735–7740. https://doi.org/10.1073/pnas.1802537115 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1802537115

[167] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella M. Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *The 2014 ACM Conference on Ubiquitous Computing, UbiComp '14, Seattle, WA, USA, September 13-17, 2014*, A. J. Brush, Adrian Friday, Julie A. Kientz, James Scott, and Junehwa Song (Eds.). ACM, 3–14. https://doi.org/10.1145/2632048.2632054

[168] Xu Wang, Zimu Zhou, Fu Xiao, Kai Xing, Zheng Yang, Yunhao Liu, and Chunyi Peng. 2019. Spatio-Temporal Analysis and Prediction of Cellular Traffic in Metropolis. *IEEE Trans. Mob. Comput.* 18, 9 (2019), 2190–2202. https://doi.org/10.1109/TMC.2018.2870135

[169] Amy Wesolowski, Nathan Eagle, Andrew J Tatem, David L Smith, Abdisalan M Noor, Robert W Snow, and Caroline O Buckee. 2012. Quantifying the impact of human mobility on malaria. *Science* 338, 6104 (2012), 267–270.

[170] Robert West, Ryen W. White, and Eric Horvitz. 2013. Here and there: goals, activities, and predictions about location from geotagged queries. In *The 36th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '13, Dublin, Ireland - July 28 - August 01, 2013*, Gareth J. F. Jones, Paraic Sheridan, Diane Kelly, Maarten de Rijke, and Tetsuya Sakai (Eds.). ACM, 817–820. https://doi.org/10.1145/2484028.2484125

[171] Matthew J. Williams, Roger M. Whitaker, and Stuart M. Allen. 2012. Measuring Individual Regularity in Human Visiting Patterns. In *2012 International Conference on Privacy, Security, Risk and Trust, PASSAT 2012, and 2012 International Confernece on Social Computing, SocialCom 2012, Amsterdam, Netherlands, September 3-5, 2012*. IEEE Computer Society, 117–122. https://doi.org/10.1109/SocialCom-PASSAT.2012.93

[172] Brian M Wood, Jacob A Harris, David A Raichlen, Herman Pontzer, Katherine Sayre, Amelia Sancilio, Colette Berbesque, Alyssa N Crittenden, Audax Mabulla, Richard McElreath, et al. 2021. Gendered movement ecology and landscape use in Hadza hunter-gatherers. *Nature human behaviour* 5, 4 (2021), 436–446.

[173] Fengli Xu, Zhen Tu, Yong Li, Pengyu Zhang, Xiaoming Fu, and Depeng Jin. 2017. Trajectory Recovery From Ash: User Privacy Is NOT Preserved in Aggregated Mobility Data. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, Rick Barrett, Rick Cummings, Eugene Agichtein, and Evgeniy Gabrilovich (Eds.). ACM, 1241–1250. https://doi.org/10.1145/3038912.3052620

[174] Yang Xu, Alexander Belyi, Iva Bojic, and Carlo Ratti. 2018. Human mobility and socioeconomic status: Analysis of Singapore and Boston. *Comput. Environ. Urban Syst.* 72 (2018), 51–67. https://doi.org/10.1016/j.compenvurbsys.2018.04.001

[175] Takahiro Yabe, Kota Tsubouchi, Naoya Fujiwara, Takayuki Wada, Yoshihide Sekimoto, and Satish V Ukkusuri. 2020. Non-compulsory measures sufficiently reduced human mobility in Tokyo during the COVID-19 epidemic. *Scientific reports* 10, 1 (2020), 1–9.

[176] Takahiro Yabe, Kota Tsubouchi, Yoshihide Sekimoto, and Satish V. Ukkusuri. 2022. Early warning of COVID-19 hotspots using human mobility and web search query data. *Comput. Environ. Urban Syst.* 92 (2022), 101747. https://doi.org/10.1016/j.compenvurbsys.2021.101747

[177] Takahiro Yabe, Kota Tsubouchi, Yoshihide Sekimoto, and Satish V. Ukkusuri. 2022. Early warning of COVID-19 hotspots using human mobility and web search query data. *Comput. Environ. Urban Syst.* 92 (2022), 101747. https://doi.org/10.1016/j.compenvurbsys.2021.101747

[178] Jane Yakowitz. 2011. Tragedy of the data commons. *Harv. JL & Tech.* 25 (2011), 1.

[179] Nicholas Jing Yuan, Yingzi Wang, Fuzheng Zhang, Xing Xie, and Guangzhong Sun. 2013. Reconstructing Individual Mobility from Smart Card Transactions: A Space Alignment Approach. In *2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, December 7-10, 2013*, Hui Xiong, George Karypis, Bhavani M. Thuraisingham, Diane J. Cook, and Xindong Wu (Eds.). IEEE Computer Society, 877–886. https://doi.org/10.1109/ICDM.2013.37

[180] Sihan Zeng, Huandong Wang, Yong Li, and Depeng Jin. 2017. Predictability and Prediction of Human Mobility Based on Application-Collected Location Data. In *14th IEEE International Conference on Mobile Ad Hoc and Sensor Systems, MASS 2017, Orlando, FL, USA, October 22-25, 2017*. IEEE Computer Society, 28–36. https://doi.org/10.1109/MASS.2017.32

[181] Desheng Zhang, Jun Huang, Ye Li, Fan Zhang, Cheng-Zhong Xu, and Tian He. 2014. Exploring human mobility with multi-source data at extremely large metropolitan scales. In *The 20th Annual International Conference on Mobile Computing and Networking, MobiCom'14, Maui, HI, USA, September 7-11, 2014*, Sung-Ju Lee, Ashutosh Sabharwal, and Prasun Sinha (Eds.). ACM, 201–212. https://doi.org/10.1145/2639108.2639116

A ETHICS

Not applicable.