Department of Computer Science

Temporal patterns of human behavior

Talayeh Aledavood

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 225/2017

Temporal patterns of human behavior

Talayeh Aledavood

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall T2 of the school on 5 December 2017 at 12 o'clock noon.

Aalto University
School of Science
Department of Computer Science
Complex Systems

Supervising professor
Jari Saramäki

Preliminary examiners
Alain Barrat
Centre de Physique Théorique
Marseille, France

Joachim Mathiesen
Niels Bohr Institute
University of Copenhagen, Denmark

Opponents
Ciro Cattuto
ISI Foundation
Turin, Italy

Aalto University publication series
DOCTORAL DISSERTATIONS 225/2017

ISBN 978-952-60-7723-9 (printed)
ISBN 978-952-60-7724-6 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-7724-6

Images: Cover image: National Photo Company (1924). Bonus Bureau, Computing Divison [photograph]. Library of Congress Prints and Photographs Division, Gift; Herbert A. French; 1947. Retrieved from http://www.loc.gov/pictures/item/npc2007012636/

Unigrafia Oy
Helsinki 2017

Finland

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Talayeh Aledavood

Name of the doctoral dissertation
Temporal patterns of human behavior

Publisher School of Science

Unit Department of Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 225/2017

Field of research Computational Science

Manuscript submitted 21 August 2017 **Date of the defence** 5 December 2017

Permission to publish granted (date) 4 October 2017 **Language** English

☐ **Monograph** ☒ **Article dissertation** ☐ **Essay dissertation**

Abstract

With the development of programmable computers, humans have entered the digital age. The emergence of the World Wide Web and the ubiquity of computers, mobile phones, and other devices that automatically store digital records have led to the concept of *big data*. To harness this big data, new computational tools and methods are constantly being created to extract information from it. When people interact with digital devices and platforms, they leave digital footprints. These traces can open a window into understanding the behavioral patterns of humans. The emerging multi-disciplinary field of computational social science takes advantage of the large, empirical datasets built of these footprints and uses them to address questions from various fields of social sciences by applying methods and techniques from hard sciences like physics and network science. In the past decade, there has been a surge of studies where such datasets have been used to study human patterns of behavior. Many have looked at structural properties of social networks such as personal network sizes or tie strengths. A more recent trend focuses on temporal features of human behavior and communication. In this thesis, multiple datasets of digital activity have been analyzed. These data are of various types, from communication timestamps to sociodemographic data.

The main focus of this thesis is to understand temporal patterns of human behavior, such as daily and weekly patterns of communication, as well as patterns of mobile phone usage, which can be seen as proxies of times of sleep and wakefulness. Looking at these different rhythms, we find that individuals exhibit activity patterns which are unique to each person and they tend to maintain their signature activity pattern over time.

Based on their propensity to sleep at different hours of the day, people can be categorized into groups called *chronotypes*. By analyzing the phone usage activity, using a dataset from a study with 1000 participants, we infer individuals' chronotypes and find that people with different chronotypes vary in the features of their personal social network, such as the number of their contacts. For example, we see that evening-active individuals maintain larger networks. Also, by looking at the social network of study participants we observe that evening-active people tend to be more central in the network. They also exhibit homophily, which is absent for morning-active individuals.

Recently, much effort has been made to design studies which combine different devices and data sources to collect data from individuals with the goal of addressing specific questions and trying to tackle societal challenges such as the spread of diseases or issues of mental health. We have worked in a multi-disciplinary group to design a prototype data collection platform, which is currently being used for projects ranging from mental health to neuroscience studies.

Keywords computational social science, temporal patterns, big data, social networks, data collection studies

ISBN (printed) 978-952-60-7723-9 **ISBN (pdf)** 978-952-60-7724-6

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2017

Pages 155 **urn** http://urn.fi/URN:ISBN:978-952-60-7724-6

For Mansooreh & Ali,
who gave me the gift of being.

Preface

First, this is not one of those acknowledgements which starts with “First and foremost”. Like everything else, I like my acknowledgements organized chronologically. Because this is perhaps my last chance to write an acknowledgement which can be as long as I want, it is going to be a long list of people that I would like to thank: those who contributed to my education and passion for science before I came to Finland. This is by no means comprehensive, but rather a list of those who have been truly inspirational to me and have believed in me.

First, my parents, because they were the first people I have known, as well as my first educators. As long as I can remember, my mom was reading books to me, and when I asked for more and more she would still read more. Although I did not notice this until much later in life, simply observing my dad work has taught me how research is done: lots of hard work, curiosity, and perseverance. As far back as I can remember, my dad was either reading, writing, or answering my endless questions. My mom has also taught me to never stop learning. I recall how she was always reading and preparing before her lectures to stay up-to-date, even after 30 years of teaching.

My grandmother "Mamani Hajieh", who was immeasurably kind, always told me stories and answered my questions. She was always patient and valued education perhaps above everything else. She was my first role model in life.

My uncles Mohammad Reza and Amir Ahmad, who have always inspired me (each of them in a different way) and have always believed in and continuously encouraged me. My aunt Manijeh, from whom I have learned to never stop and to never let anything stop me. I have also learned from her to always welcome challenges and find a way to solve them. She is my second role model in life.

My cousin Mahsa, whom I would rather call my sister, who has been teaching me since I was born, and still does.

My brother Parham, who was in a way my first student. I used to share my "wisdom" with him and teach him what I knew. I enjoyed it very much because he was always a quick learner. Even though he didn't believe that "Avalanche" is a good book that he must read–and only reading it much later found to no surprise that it is a great book–he has always allowed me to advise him and to share my knowledge with him. This however was mainly the case when he was a kid, and nowadays it is me who learns from him all the time. I should also note that he has the best sense of humor one might imagine.

My first grade teacher Ms. Sabouri, my third grade teacher Ms. Jalili (here I cheated a bit because going chronologically Parham should have been somewhere here, because he was born in the middle of my third grade), and my fifth grade teacher Ms. Mirani, who always encouraged me to be better (for some strange reason I had a better relationship with teachers in odd grades).

In middle school, two of my physics teachers, Ms. Sharikzadeh and Ms. Erissian, who have both had an important role in my decision to study physics later on. I should also especially thank Ms. Erissian for her nice signature, whose style I used to create my signature (which I still use). Ms. Mokhtari, Ms. Ahani, and Ms. Ziaolmolki, who were all great educators that taught me many important lessons. Ms. Farid Moayer, who taught me to make crafts and artwork that I still sometimes make.

Haleh Olfati, who was my first mentee and made me really proud when she gave her presentation on Aristotle at the school's auditorium in front of a large crowd visiting the school fair.

Nazanin Hassannia and Golnoosh Bizhani, who shared the passion of the sky with me and who taught me a lot related to constellations and sky observations.

Mr. Torabi, my high school combinatorics teacher, who re-sparked my joy in math.

Reza Mansouri, one of my professors at Sharif University, who trusted me and encouraged me to attend an international conference.

Jan Friedrich, at TU-Munich, who spent many hours teaching me how things work. His passion for science and research was truly inspiring and kept me motivated as a researcher. He also gave me the opportunity to work at CERN several times. I have had invaluable experiences because

of his help.

My friends Sheema and Elham, who have always believed in me and helped me in many ways. There is no way I can thank them, so I will just make sure to never lose them as friends.

Mélina, who supported me and encouraged me when I needed it most.

With this, I am almost at the time when I moved to Finland and started my doctoral studies. I was lucky to get support and encouragement from my friend Taha for applying to the PhD position available in the Complex Systems group at Aalto University. I first came to Finland on a cold and dark November night in 2012. My arrival was at night, so there was no wonder that it was dark out. However, there still wasn't much improvement in the light level until a few months later when the sun appeared again. To be honest, at first I wasn't sure if I can make it in cold and dark Finland, but very soon everything started to get better. The reason for this was the help and support of my great colleagues. Only a week after my arrival they all came to help me to assemble my IKEA furniture. After that, there were years of brunches, movie nights, new year parties, and most importantly, interesting research. If it wasn't for their help I would not have been able to do this and to make Finland my home. Darko, Richard, Juan, Marija, Arnab, Hang, thank you for always being there. I would also like to thank Raj, Lauri and Gerardo from whom I have learned a lot at different times. My other colleagues and office mates: Onerva, Ilkka, Rainer, Pietro, Tuomas, Ana, Sara, Arash, and Christopher, thank you for all these years, all the good chats, coffees, and everything else.

I was very fortunate to visit three research groups outside of Finland during my PhD time: Sune Lehmann's group at DTU, JP Onnela's lab at the Harvard T.H. Chan School of Public Health, and Taha Yasseri's group at the Oxford Internet Institute. I am grateful to all of them for giving me these great opportunities. I enjoyed all my visits and learned a lot from from them.

I would like to thank John Torous, whom I met during my visit to Harvard, who taught me a lot about digital mental health. Also, by observing his working style, I learned how to work more efficiently.

Many thanks goes to Kimmo Kaski, who has always provided interesting discussions. He has given support and helped me achieve my goals.

I am grateful to Erkki Isometsä and Jesper Ekelund for their support in forming a very promising collaboration in digital mental health. I am planning to pursue this line of research after graduating, and look forward

to many more years of fruitful collaboration.

I would like to thank Jussi Autere, the head of EIT digital doctoral school in Helsinki, who has always helped me with all the issues related to the doctoral school or anything related to entrepreneurship in general. I would also like to thank Katri Sarkio at EIT digital for her support throughout the years.

Special thanks goes to all my coauthors, who have made the completion of this PhD work possible: Robin Dunbar, Sam Roberts, Esteban Moro, Felix Reed-Tsochas, Eduardo López, Rainer Kujala, Sune Lehmann, Ana Triana Hoyos, Tuomas Alakörkkö, Kimmo Kaski, Erkki Isometsä, Richard Darst, and Jari Saramäki.

I am thankful to Leo Kärkkäinen, my supervisor during my internship at Nokia Digital Health Lab, and his wife Asta, who have both been inspirations to me. I am very happy that I have been able to meet and know them.

Uli, Shirin, Daniel, Lara, Jussi, Anna, D&D, J&J, Parisa, and Tolou: thank you for all the friendship, support, games, dinners, lunches and brunches.

Mikko (Kivelä), thank you for being a good friend as well as a great mentor and scientist.

I would like to thank Ciro Cattuto, who has agreed to act as my opponent and (I am sure) will provide a lively discussion during my defense. I am also grateful to Alain Barrat and Joachim Mathiesen, my pre-examiners, who provided encouraging and helpful comments for this thesis.

I am filled with gratitude to my advisor Jari Saramäki, who is not just a great scientist, but also a great human being. I have learned so much from him and am grateful for so many things that I decided to write to him separately in length, to tell him about all the things I have learned during the past five years and am thankful for.

Nima, thank you for always sleeping there with the occasional meows. You keep me going! In the end, I would like to thank Richard, without whom everything in life is just less good. You are simply amazing.

Helsinki, November 14, 2017,

Talayeh Aledavood

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Talayeh Aledavood, Sune Lehmann, Jari Saramäki. Digital daily cycles of individuals. *Front. Phys.*, Volume 3, p. 73, October 2015 .

II Talayeh Aledavood, Eduardo López, Sam G. B. Roberts, Felix Reed-Tsochas, Esteban Moro, Robin I. M. Dunbar, Jari Saramäki. Daily Rhythms in Mobile Telephone Communication. *PLOS ONE*, Volume 10, Issue 9, e0138098, September 2015 .

III Talayeh Aledavood, Eduardo López, Sam G. B. Roberts, Felix Reed-Tsochas, Esteban Moro, Robin I. M. Dunbar, Jari Saramäki. Channel-specific daily patterns in mobile phone communication. *Proceedings of ECCS 2014*, Springer International Publishing, p. 209–218, May 2016 .

IV Rainer Kujala, Talayeh Aledavood, Jari Saramäki. Estimation and monitoring of city-to-city travel times using call detail records. *EPJ Data Science*, Volume 5, p. 6, March 2016 .

V Talayeh Aledavood, Sune Lehmann, Jari Saramäki. Social Network Differences of Chronotypes Identified from Mobile Phone Data. *submitted*, 8 pages, August 2017 .

VI Talayeh Aledavood, Ana Triana Hoyos, Tuomas Alakörkkö, Kimmo Kaski, Jari Saramäki, Erkki Isometsä, Richard K. Darst. Data Col-

lection for Mental Health Studies Through Digital Platforms: Requirements and Design of a Prototype. *JMIR Research Protocols*, Volume 6, Issue 6, e110, June 2017 .

Author's Contribution

Publication I: "Digital daily cycles of individuals"

Major role in developing the ideas, designing the research, implementing the methods, and analyzing the data. Primary writer.

Publication II: "Daily Rhythms in Mobile Telephone Communication"

Major role in designing the research, implementing the methods, and analyzing the data. Helped with writing the article.

Publication III: "Channel-specific daily patterns in mobile phone communication"

Major role in developing the ideas, designing the research, implementing the methods, and analyzing the data. Primary writer.

Publication IV: "Estimation and monitoring of city-to-city travel times using call detail records"

Helped with developing the ideas, designing the research, implementing the methods, and writing the article.

Publication V: "Social Network Differences of Chronotypes Identified from Mobile Phone Data"

Original concept of the research. Major role in designing the research, implementing the methods, and analyzing the data. Primary writer.

Publication VI: "Data Collection for Mental Health Studies Through Digital Platforms: Requirements and Design of a Prototype"

Original concept of the research. Major role in designing the research, running the experiments, and writing the article.

List of Figures

1. The Digital Age

Computers, data, and computation are focal points of modern humans' lives. Even though they all have existed in different forms since long ago, the rates of change has been getting faster and faster. In the late 19^{th} century, the advent of electrification [1] and the spread of the first electronic communication technology (the telegraph) resulted in the second industrial revolution [2]. In the mid-20^{th} century, computers and automated data processing began in earnest as part of the technological revolution. Now, we are going through the next great transformation: the ubiquity of information, and again are experiencing irreversible changes in human society [3, 4].

Electronic data processing was born for certain limited tasks, specifically cryptanalysis during World War II [5]. Now, data processing is a critical skill for modern society. Effective use of data allows even more data to be created, causing an exponential growth in the total information available. As scientists, we can use this data to study and learn about society [6].

1.1 Computers and computational power

Although it may surprise us now, technology for computation has existed long before electronic computers. Since thousands of years ago, humans have developed techniques and built devices for computation: counting, abacuses, arithmetic, and slide rules are all examples of manual computation methods [7–9]. However, computational devices completely transformed the human society in the 20^{th} century when they became fully automated [5]. The first programmable and fully digital computer was built in 1941. Six years later, the first transistor was made [10]. Since the 1970's, the number of transistors contained in central processing units

has famously grown exponentially (Moore's law) [11]. Along with this, the size of computers has become smaller and smaller and the energy requirements have decreased dramatically.

In the past decades the form of computers has changed from massive machines only operated by experts for certain computationally oriented purposes to devices that have made their way to our households and even to our pockets and our wrists. The ubiquity of computers does not only mean that their scope of usefulness has expanded, but also that there are vast amounts of data being produced as a result of their usage.

The most recent revolution is not just in power, but in size. Smaller sizes have allowed computers to become pervasive—currently, in developed countries, there are more mobile phones than people, and even in developing countries the penetration rates are more than 90% [12]. Smart phones, the newest generation of mobile phones, have also become extremely popular—at least in some parts of the world—especially after Apple introduced iPhones in 2007. Smartphone users depend on them for many everyday activities, such as navigation, communication, listening to music, searching for information, taking pictures, and much more. In order to provide such a wide range of functions, smartphones contain a multitude of sensors, from accelerometers to GPS and light sensors.

The original purpose behind the advances in mobile phone technology and hand-held devices was their utility and the fact that they make interactions easier, people more connected, and information better available. However, as a result of interaction of people with these devices, digital data on humans are being produced and stored at a staggering rate [13]. A substantial part of these data can be seen as behavioral information [14].

For scientists, these pervasive computers serve primarily as sensors that collect more and more data for their research. They can also serve as a way to interact with subjects, as we will see in Chapter 4.

1.2 Data

The continuous production of large amounts of digital data (the so-called data deluge [15]) is a consequence and signature of the digital age. It is not only the computational power which has advanced drastically, but also the amount of data which can be (and is) produced, stored and transmitted. *Big data* is a term which is often used to address this new data production. The term is somewhat of a marketing term, but it is usu-

ally described in terms of volume, velocity, and variety: large amounts of data, data arriving and being processed at a high rate, and many different formats such that it is difficult to handle data uniformly [16, 17]. Scientifically, experimental physicists were some of the first to collect huge data volumes, for example at CERN (European Organization for Nuclear Research), and they continue to be some of the largest producers of scientific data [18].

CERN has had another significant role in the digital age. The World Wide Web (WWW or simply the web) which has become a major part of our lives today, was first invented at CERN in 1989 [19, 20]. One of the reasons behind the success and expansion of the web is that it revolutionized the way people can connect to each other and makes interactions very easy. This has led to the rise of today's so-called *networked society*, in which society is centered around electronic devices and information which circulates through them [21]. The use of the web and digital devices have become extremely pervasive, and lead to the production of vast amounts of digital data and traces that people leave behind as a result of online interactions [22]. Also, electronic devices are becoming smaller and smaller everyday, constantly increasing the data production rate. For example, in 2017, 40 billion text messages were sent only in India on New Year's Eve through WhatsApp messaging platform [23] and in 2016 there were 187 million tweets about the summer Olympics in Rio de Janeiro [24].

The advent of small embedded computers allows data to come not just from single devices, but from a wide range of sensors across different devices. These sensors allow large amounts of data to be collected, both from specific people and from general environments. For example, in addition to mobile phones, the use of other wearable devices such as activity trackers is becoming increasingly popular [25]. These devices have sensors which can measure different types of movement as well as physiological parameters such as pulse rates [26]. Combining this with data from other devices can give us detailed, individual-level behavioral data on human behavior [27].

In order to make sense of these data, a multitude of tools have been created for data handling. Many of the tools or concepts were developed in companies, and are now easily available as parts of various open source projects [28]. As just one of many examples, the Apache Hadoop Distributed Filesystem (HDFS) is essentially a distributed, replicating storage system which can be used to store data on the scale of petabytes

and larger [29, 30]. HDFS integrates with Apache Hadoop, which provides a framework for batch computation [31, 32]. The Hadoop ecosystem also includes a wide variety of tools for processing and transferring data, for example Apache Spark for higher-performance analytics and Apache Kafka for data connection and redirection [33–35]. Hadoop is just one of many available frameworks. In addition to having software available, entire platforms are available as a service. For nothing other than money, one can buy ready-made platforms in the cloud that can readily use any of these technologies and scale to any size of data [36]. Thus, the most pressing problem is not the availability of methods for data handling, but how to use the data.

1.3 Opportunities and human benefits

In a paper published in 1996 [37], Fayyad *et. al.* write: "Across a wide variety of fields, data are being collected and accumulated at a dramatic pace. There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data". Even though the paper is more than two decades old, this statement remains true even today. Despite the fact that computational capabilities in terms of tools, techniques, and theories have shown major advances, the digital data production rate has also kept growing. This has brought scientists across many disciplines valuable research opportunities as well as many challenges. Ever since computers were invented, it has been known that it is not enough that computers can do computations faster than humans. We must also "teach" them to find patterns in the data in the same ways that humans can [38]. This is why fields such as machine learning and artificial intelligence have emerged, where machines are taught to perform tasks such as recognizing handwriting and speech, driving vehicles, or replicating tasks that medical experts do [38]. Another powerful tool which is being used by a growing number of researchers since the late $1990'$s is *network science*, which is discussed in more detail in Chapter 2.

In the remaining chapters of this thesis, we will describe how to make sense of auto-recorded data on individuals using the emerging field of computational social science and discuss the challenges and opportunities within this field. We show how digital data from modern electronic devices and platforms can provide rich behavioral information on humans.

We also look at temporal patterns of different human behaviors such as communication, sleep, and rest, both at the collective level as well as at the individual level through the analysis of such data. We show how we can scale up studies of humans by using new methods of computational science and big data, instead of using traditional methods of data collection from humans such as surveys. In the final chapter, we discuss design of data collection platforms for controlled studies which are specifically designed to collect certain types of digital data from human subjects with pre-defined research goals and questions.

2. Computational Social Science

The digital age has one important effect in the way scientific work is done: computation has become an important part of the modern day's scientific endeavors [39]. Computational science takes advantage of computers to do calculations much faster and at much larger scales and speed than possible by humans. While the dominance of computational science has already begun decades ago in the natural sciences, such as physics or biology, the process has been much slower for the social sciences [40]. Computer simulations of social systems and agent-based modeling were perhaps among the first use cases of computers in social sciences. The genesis of such simulations dates back to the 1960's, with a leap of attention in the 1990's [41]. However, the use of big data in social sciences did not start until about a decade later.

Data, which is the fuel of quantitative social sciences, is traditionally gathered by means of surveys (*e.g.* by interviewing the study participants by phone or in person), questionnaires, and observational methods [42, 43]. This makes data collection extremely costly and labor intensive. However, in today's world there is an abundance of digital data on human behavior from various sources and devices, which can be used for social science studies. Modern computational power and advanced statistical methods applied to these big data enable us to study social systems as well as individual behavioral patterns.

In the modern world, uncovering human behavioral patterns and understanding the governing mechanisms of human interactions is key to addressing big societal challenges [44]. Within the past few decades, issues such as the spreading of diseases [45, 46], data security and privacy [47, 48], and urban planning and traffic management [49, 50] have increasingly been studied within multidisciplinary groups of researchers harnessing big data and using new computational methods.

The field of computational social science (CSS) is an emerging field which tries to address different questions within social sciences with the use of modern computational methods and big data. CSS brings together scientists from a wide range of disciplines within social sciences, such as sociology, economy, and psychology, as well as hard sciences from physics to computer science and statistics. Even though CCS is quickly evolving, it has had a rather slow start. In this chapter, we discuss some of the challenges which have impeded the progress of CSS both in the past and at present as well as data sources for CSS studies. Finally, we introduce network science which has been used as an important tool in these studies.

2.1 Challenges in CSS

Some of the challenges described here were initial obstacles for the formation of CCS as a discipline. However, others are ongoing challenges which exist as a result of the complexities of data, methods, or collaborations between disciplines. Below, we discuss 5 key challenges for CSS studies.

1. The existence of digital traces of human behavior which form large empirical datasets does not necessarily mean that the data are available to be accessed by researchers. Technology giants such as Facebook [51], Google [52], and telecommunication companies have access to vast amount of data on individuals' activity, behavior, characteristics, and communication. They can (and do) analyze this data to gain knowledge about individuals for the purpose of business intelligence. However, these data and the gained knowledge are often treated as proprietary. Advances in CSS have become possible only if such data have become available for anyone to download, or if they have been shared with a group of researchers under nondisclosure agreements (NDA). Examples of both types of studies are discussed further in this chapter in Section 2.2. Another way to get access to data is for scientists to run their own data collection studies with the purpose of collecting digital data from individuals to address specific research questions. Such studies are discussed in detail in Chapter 4.

2. There is a lot of synergy between the natural sciences and the components of digital age. For example, in the case of artificial neural net-

works, knowledge on the brain structure is used as an analogy to build an adaptive statistical model [53]. In studying neural networks, computation is so intertwined with biology that scientists have wondered: "Are we using computation as an aid in understanding biology (*e.g.* evolution, thinking, etc.) or are we using biology as a metaphor to work on computation?" [54]. The other example is the invention of the web to facilitate the work of scientists at the world's largest physics lab (CERN). However, for social sciences, this synergy has traditionally been lacking, and the integration of the computational world view of the natural sciences with social science remains a challenge.

3. It is essential for social scientists to collaborate with other disciplines to harness what the digital age has to offer. Even though very complex problems in science often can only be solved if researchers across different disciplines work together, such collaborations are often very slow and inefficient. This is for example because of the differences in the methodologies and scientific jargon, or in the aspects of the problem that each discipline finds most relevant. It is important to note that the need for collaboration exists for both sides. Social sciences provide the motivation for research and some analytical methodology developed specifically to handle social systems, while experts in computational sciences provide tools and methods to collect, manage and analyze large amounts of data.

4. In CSS, digital data are used as proxies of human behavior [44]. For example, in a CSS study about "friendship", researchers might use data from Facebook and interpret a link between two people (the virtual friendship) as friendship between the two individuals. However, on Facebook, becoming a friend with someone might only be a way of having the possibility of contacting them in the future (similar to exchanging business cards). This can lead to making false interpretations regarding the research question. But, biases are part of any kind of quantitative research, and they can be minimized with enough vigilance in processing the data and interpreting the results. In our example, one possible way to make the interpretations more meaningful, would be to filter out links (friendships) which do not have any interaction other than the original formation of the link. In CSS research, it is therefore essential to understand what is measured, what is inferred, and how one is only

a proxy of the other. Failure to do this can lead to erroneous results.

5. In large datasets missing or inaccurate data is often a source of problems. Statisticians are constantly developing methods to deal with this issue. In CSS, the problem can get even more complicated when there are multiple sources of data. Data might come from various devices, different datasets, or even be a combination of large empirical datasets and questionnaires. In addition to the problem of missing data, matching different data sources can be a burdensome task. For example, in case of temporal data, different data sources can vary in sampling frequencies, or spatial data, can have different spatial resolution or non-identical unit shapes.

Despite all the challenges described above, the field of CSS has now come into existence and it has made significant advances over the last decade, some of which are discussed in this work. In the next section, we will talk about some of the main sources of big data used in CSS studies.

2.2 Sources of Big Data

2.2.1 Social networking websites and apps

Facebook recently announced that the number of its monthly users has reached 2 billion [55]. This means that more than a quarter of world's population who are connected in the real world are also connected to each other on this virtual platform. Their social network, their public profile, as well as all their on-platform interactions are registered with Facebook. This makes a database far larger and richer than what any usual social-science study which collects data by surveying users can ever create. Facebook is not the only example of such social networks: there are many other platforms, such as Twitter [56], Instagram [57], or LinkedIn [58], where users produce content and interact with others. Depending on the nature of the platform, the links between two individuals indicate different things such as friendship (*e.g.* on Facebook) or following (*e.g.* on Twitter.) Thus, analyzing data from different platforms can serve to answer different questions. Even though these websites have rich data in abundance, this does not always mean that the data are available to researchers.

However, sometimes the companies behind these websites have their own research and development teams that analyze the influx of data. There have even been cases where companies such as Facebook have run social experiments through their website, leading to much controversy and discussions with respect to principles of informed consent and right of users to opt out of such experiments [59, 60].

One possible way to use the data produced by these websites is public APIs (application program interfaces). APIs are interfaces by which data from one platform can be pragmatically accessed by something outside. The amount and scope of data which can be accessed by a third party varies from one platform to another. In some cases researchers can also apply for extra permissions which have to be examined before extra data extraction privileges are granted. An API can be a very convenient tool for a researcher, but the data owners (companies) are often equally restrictive in what can be accessed and used.

2.2.2 Wikis

A wiki is a web interface where information and knowledge is collected by collective input from various users. One of the most well-known uses is the giant encyclopedias of crowd-sourced information. The main difference between a wiki and an encyclopedia is that wikis are a collection of knowledge from all users, and almost any person can add content to them or remove them. Because of this fundamental property, wikis can be very dynamic and evolve quickly over time. The most significant wiki is Wikipedia [61], which has been created by contribution of millions of people. In the recent years, there has been a vast amount of research on Wikipedia, focusing on a wide range of social phenomena. For example, temporal activity patters [62], opinion dynamics and conflicts [63], and predicting significant social phenomena [64, 65] have all been inferred from the Wikipedia edit history. Wikis have an added advantage in their openness: all data is commonly available, and research is generally encouraged as it is seen to produce societal benefits.

2.2.3 Mobile phone Call Detail Records

The first handheld mobile phones were built in the 1970's. Mobile phones however became ubiquitous in the 21^{st} century. In 2014, 96% of the world population owned mobile phones, with penetration rates of more than

100% in developed countries [12]. Today's smartphones contain a multitude of sensors and have relatively high computing power. Thus, people who carry and use mobile phones are highly monitored, with lots of variables such as their movement and social interactions recorded passively with almost no interruption. However, even the simplest and oldest of mobile phones can still provide rich data on users. Telecommunication operators record each user's usage diary for the purpose of billing them for the provided service. These data, referred to in the industry as Call Detail Records (CDRs), usually contain the user phone number or ID, the timestamp of each communication, the communication type (a call, a text message, or something else), and the number or ID of the other party they are in contact with. Operators also record location data, typically the ID of the cell tower which the mobile phone has been connected to during the communication event. The exact location of all cell towers is known, so an approximate location of the user can be estimated for each communication instance. In addition to all this, telecommunication operators often have other types of data on their users, such as their age, gender and home address. In combination, these datasets contain lots of information about individuals. Similar to data from social network platforms, CDRs can be either used for research and marketing purposes by telecommunication companies, or shared with external researchers through NDAs. In the past decade, such data have been used extensively and they have been proven very informative in a wide range of studies. In [12], Blondel *et. al.* provide a thorough review of research in this field. Some examples include the study of social network structure and properties [66, 67], spreading processes on networks [68, 69], or people's mobility patterns [70, 71].

In the past years, large telecommunication companies such as Orange and Telecom Italia have organized data challenges in which they have provided some CDR data to researchers for analysis. The focus of such challenges is often research which can readily be applied in the real world for example in the context of health, urban planning or agriculture. In publication IV, we use CDR data provided within one such data challenge [72] to study inter-city travel times in Senegal. The provided dataset contained anonymized communication records of Orange subscribers in Senegal. Our goal was to use CDR data to estimate travel times between cities, first because such information is often hard to find or inaccurate in developing countries, and second because this would allow almost-real-time

monitoring of travel times. For different pairs of cities in the country, we identified individuals who had a registered CDR event first in some city A and later in another city B. We take this inter-event time as a proxy of the travel times between cities A and B. Note that CDRs are typically only recorded at times of communication events such as calls, and these do not necessarily coincide with arrival or departure. Therefore we need to look at large number of events, which allows us to produce a distribution of the possible (minimal) travel time between two cities—and from here produce estimations of the actual minimal travel time.

2.2.4 Data collection studies

Even though the sources mentioned above give us data on a large number of individuals, they often lack in-depth data on each person. This limits the scope of research questions that can be addressed. We cannot, for example, know how characteristics such as health, education or income of individuals relate to social network features of the person.To overcome this issue, many researchers have started to run controlled experiments where they can collect data from individuals for a certain purpose. Digital data collected in these studies are often augmented with traditional data collection methods like surveys and questionnaires. Publications I, II, III, and V use data from such experiments (made available to the research in this thesis, but collected by other research groups), and publication VI is dedicated to the design of such data collection experiments (discussed in more detail in chapter 4). Below, we describe some datasets used in this thesis:

Reality Mining

The pioneer of such data collection efforts is perhaps the “Reality Mining” study which was run at MIT in 2004 (over the course of 9 months) using custom-developed technology in Finland at the at the University of Helsinki [73, 74]. This study collected communication data as well as Bluetooth and location data from 100 participants. The data are publicly available to anyone with a minimal privacy agreement [75]. A part of this dataset has been used for this thesis in publication I.

UK students dataset

This experiment, was carried out in the UK in $2007 - 2008$ [76]. In this study, mobile phones were distributed to 30 students in the last year of their secondary school. They had the phones for 18 months and they filled out extensive questionnaires about all people in their social networks every six months (three times total). Access to this private dataset was obtained through collaborations with the researchers at the University of Oxford and the University of Chester who collected the data. Data from this study were analyzed in publications II and III.

Copenhagen Networks Study

In this study which started in 2013, about 1000 incoming students at the Technical University of Denmark (DTU) were given identical smartphones. Each phone came with an application which was designed to collect data from various sensors of the phone, such as Bluetooth, mobile phone screen activity, location, and communication events. All participants also filled out several questionnaires about their health and psychological profile. The study is described in detail in [77]. This private dataset was accessed through collaborators at the Technical University of Denmark. Data from this study are used in publication V.

2.3 Network science: a powerful tool for CSS

CSS is a truly multi-disciplinary field. As a result, the list of tools and methods applied in this field is long and constantly growing. Here, we focus on one specific tool which has proven to be useful in many studies: *network science*. Most social systems can be modeled as networks, where nodes are individuals and links are interactions or affiliations between them. The origin of network science is in graph theory, a branch of mathematics which roots back to the 18th century [78]. The famous problem of the "Seven bridges of Königsberg" was solved in 1735 by the Swiss mathematician Leonhard Euler and marked the beginning of this field [79]. The use of graph theory in sociology, commonly referred to as *social network analysis* started in the 20^{th} century, with growing interest in the topic from the researchers in the 1970's [80]. However, the network science subfield of complex systems originated in the work of complexity

scientists only about two decades ago. Complex networks were introduced in the works of Watts and Strogatz [81], and Barabási and Albert [82].

Social networks, which have been widely studied by network scientists and sociologists, have been shown to have many shared characteristics. An important property which is commonly defined for all types of networks is the *degree*. The degree of a node is the number of nodes which are connected to it [80]. A characteristic which is often subject to interest in networks is the distribution of node degrees. Social networks often have broad, fat-tailed degree distributions [83]. In addition to that, most social networks share another feature: they are highly clustered, meaning that there is a high number of triangles [84]. For example in a friendship network, this would translate to friends of one individual also being friends with one another.

The node *centrality* is an attribute of social networks which is of interest to many researchers, as it can have many practical implications. Centralities in social networks try to measure how "important" a person is within the social network and how each node contributes towards the overall structure of the network [84]. Depending on the context and use case, there are various types of centrality measures which can be defined. Examples of such measures are the eigenvector centrality, the closeness centrality, and the betweenness centrality [85].

Furthermore, we can commonly find "homophily" in social networks. Homophily refers to the property that nodes of similar characteristics tend to have links between them [86]. For example, people tend to interact with others who have similar sociodemographic properties as themselves.

Many properties of social networks can also be found in a wide variety of networks, from food webs [87] to protein-protein interaction networks [88], power grids[89], and many more [90, 91]. This means that a multitude of methods and findings which are constantly being produced can be readily applied to other types of data, for example to social networks. However, it is important to note that without having expert domain knowledge from social scientists, there is a high risk of misinterpreting the results or missing the important conclusions which can be drawn from social network data.

2.4 Summary

Modern computational tools, methods, and the emergence of social platforms as well as other sources of digital data on human dynamics have provided an unprecedented opportunity for the field of quantitative social science. By harnessing them, we can address research questions that were previously impossible to approach in social sciences. This, however, requires close collaboration between social scientists and experts in computational sciences. This thesis, which is a result of multiple such collaborations, tries to better understand temporal dynamics of human behavior through analyzing digital activity patterns.

3. Temporal Dynamics of Human Activity

"The first characteristic of modern machine civilization is its temporal regularity. From the moment of waking, the rhythm of day is punctuated by the clock."

– Lewis Mumford, *Technics and Civilization*

Time is a notion which is important in most research disciplines as well as in our everyday lives. However, what "time" means and signifies varies from one field to another. For humans, and all other living creatures on earth, following the light-dark cycle caused by the movement of Earth around the Sun is a matter of survival. Following these rhythms guarantees that there is less chance of starvation or premature death as a cause of predation [92]. In short, this means that humans tend to sleep at night, when it's dark, and are active when there is light.

Speaking of time, we can think of at least three different types: one is the *natural time*, the time which is dictated to us by the Sun and the one which humans have tried to measure with various methods and equipments to greater and greater accuracy since thousands of years ago. The other one is the *physiological time*; almost all living cells have an internal clock, which helps them to synchronize with all the other cells in their organism [93, 94], other creatures, and the natural time [95, 96]. The third one is the *social time*, a social construct, defined by sociologists, and formed as a result of collective engagement of humans with the social world [97].

In this work we try to understand how these "different times", and the temporal patterns that exist as a result of them, manifest themselves in human daily life and activity and resting times. We first study patterns within social systems, the so-called sociotemporal patterns, by analyzing digital traces that individuals within the systems leave behind. Sociotemporal patterns regulate social systems, so by uncovering them we can

learn about the properties of the system that we are studying. Temporal patterns in nature—for example temporal patterns of extraterrestrial objects—have been studied for thousands of years. Physiological temporal patterns have also been observed since ancient times and they have been studied in plants and later in humans in the past few centuries [95]. In social sciences, in fields such as economics the importance of studying temporal patterns has long been known, but this is not the case for sociotemporal patterns [98]. Émil Durkheim, commonly referred to as the father of the modern sociology, was one of the pioneers of studying patterns and regularities in the social time, as well as social events which occur with certain intervals [99, 100].

Collective temporal patterns are a superposition of individual activity patterns. Thus, in this work, after looking at patterns at the system level, we zoom in and focus on temporal patterns and regularities in individuals' activities. We first look at communication patterns and then move on to studying sleeping and resting behavior, as well as inter-individual differences in their behavioral patterns. Finally we show how different features of an individual, such as their age, gender, or sleeping preferences, affects the way they communicate with others within their social network.

3.1 Social systems through the lens of digital data

Studies of the structure of social networks, which are based on interactions (links) between humans (nodes), date back to the mid-twentieth century. This trend has continued in CSS research through the processing of large techno-social datasets [101]. CDR datasets, for example, have been extensively used to study different structural features of networks such as tie strengths [67], degree distributions [102], clusters [103], and motifs [104]. However, social interactions are extremely complex in nature. They happen through various channels, forms and media [105]: communication between two individuals can happen over the phone, face to face, by email, and so on. Also, social networks are quite dynamic. There are bondings between people which form and decay all the time [44]. This means that reducing social networks to static graphs is in many cases too simplistic. The static approach can certainly tell us about the structural properties of the system, but does not provide any details about its dynamics. To embrace the complex nature of social systems, in recent years, there has been a shift of focus from structural and static properties of

the networks towards their dynamical and temporal properties. Another reason behind this shift has roots in the scientific approach and common methodology that physicists often apply to the systems they study: modeling their behavior with the aim of uncovering the underlying mechanisms and predicting how the system evolves over time. A notable number of researchers in the field of CSS, coming from a physics background, have tried to model people's behavior in social systems by means of large empirical datasets. It should be noted that physicists attempting to understand and to model social systems is not a recent phenomenon, and the origins of the term "social physics" even date back to the 19th century [106, 107]. Ever since, physicists have tried to apply their methodology (mainly borrowed from statistical physics) to social systems [108, 109]. A common approach in studying the dynamics of human activity has been to model it as a Poisson distribution by simply assuming that human activity is Markovian [44]. This means that, for example, in a communication network, links (interactions) between two individuals are created at random points in time. Today, we know that human activity and behavioral patterns are not randomly spread in time, but rather appear in bursts of activity followed by periods of inactivity [110, 111]. This bursty behavior can be seen in individual activity patterns as well in inter-individual interactions [44].

Even though modeling human dynamics often oversimplifies matters and does not capture the essence of social systems [112], each model can be seen as a foundation for future findings. Such models bring us closer to uncovering the complexity of social systems [113], and this is exactly what the physicists' models of human behavior have been doing so far.

The "small world" phenomenon [81], also known as the "six degrees of separation", which is about the existence of short paths between individuals in social networks, is widely known even to the general public [114]. However, a rather recent study, "Small but slow world: How network topology and burstiness slow down spreading" [69] showed that despite the small world property in social networks, spreading processes (e.g. of news, rumors or disease) may slow down if tie activations between nodes (events) are bursty and inhomogeneously spread in time. This finding made it ever so clear that the inherent burstiness of human activity plays an important role in dynamical processes on networks. By aggregating ties in a network and losing the temporal features, we discard way more than we can afford if we wish to meaningfully explain dynamical pro-

cesses on social networks. This has given rise to a relatively new field in network science called "temporal networks", where the temporal features of a network are not discarded by aggregation [115, 116].

Having discussed the importance of temporal features of human activity both at the individual level as well as at the network level, we next look at temporal rhythms of different social systems.

3.2 Rhythms of activity in social systems

In social systems, while activity of individuals can vary substantially from one day to the next, the system-level behavior (*i.e.* activity patterns of a group of individuals) is more robust to individual-level variation. For example, going on a trip can significantly affect one person's activity patterns from mobility to sleep and communication with others. But, if we look at communication patterns of a whole neighborhood or city, it is less likely that they change from one day to the next (unless there is a major event). Analyzing aggregated data and looking at collective temporal activity patterns of individuals within a social system can teach us about the properties and characteristics of the system. For example, in [117], CDR data from the city of Boston is aggregated for small sections of 200 meters by 200 meters within the city, and weekly activity patterns are built for each section. The weekly rhythms in different sections are shown to correlate with the land use there. In [118], CDR data is aggregated at the city level for various cities in Spain. The authors look at daily and weekly activity patterns and see that weekends tend to have less activity compared to weekdays in all cities. The also observe that different cities vary in the activity level (which is a result of different population sizes). However, after normalization, all cities show very similar activity patterns which the authors call "urban rhythms". Looking closely into these urban rhythms, we can see slight differences between timings of activity, which is perhaps a sign of different "social culture" in different parts of the country. There are also various other studies which use different sources of digital data from urban areas to measure collective activity rhythms, which are referred to as "the pulse of the city" [119] or "urban chronotype" [120]. Studies of temporal patterns and activity rhythms in social systems are not however limited to cities. These rhythms have been studied in a wide range of systems such as online platforms like Twitter [121] or OpenStreetMap [122]. Such studies follow different purposes,

for example, understanding cross cultural differences in activity patterns of users of a certain platform [62] or uncovering temporal patterns of searching for information about certain topics [123, 124]. Publication I provides a short review of such studies. In this publication, we have also looked at communication patterns in three different social systems: the Reality Mining study (discussed in the previous chapter), a small town in a European country where we have data from two separate communication channels (calls and text messages), and finally a university for which we have timestamps of internal email communication extracted from log files of the university's mail server [125]. To build the weekly communication patterns, we divide a week into $7 \times 24 = 168$ time intervals, each of which have a width equal to one hour. For each event within the analysis range, we assign the event to the time interval it belongs to. This way we end up with a weekly activity pattern that is the result of aggregating events from all individuals within all the weeks of the study. This method assumes that there is little variation in activity for a certain hour of a given weekday from one week to another. However, to be on the safe side, we make sure that this assumption holds for the data before performing the aggregation. The outcome of this aggregation can be seen as the baseline activity rhythms of the system. Even though social systems do not show significant variations from the baseline if there are changes in the activity patterns of one or a small number of individuals, big events like large gatherings can impact many people and result in large deviations from the baseline activity. This can then be observed in digital temporal patterns of the system and even monitored in real-time, if necessary. Therefore, analyzing the temporal patterns of systems such as cities and detecting unusual behavior within them [126, 127] can also provide insights which can be used for urban planning or governing purposes.

In Fig. 3.1, we can see the aggregated weekly activity counts in 4 different cases. For example in the Reality Mining call data, all days of the week show more or less similar behavior, whereas in the email dataset the difference between the weekend and weekdays is quite pronounced. Also comparing two different communication channels within the same system, we can see that the two channels are used differently, and text messages tend to continue until later hours of the day. In summary, different systems, and even channels within the same system, exhibit different weekly rhythms, and looking at these patterns in detail and comparing them we can find information on the system in question. While looking

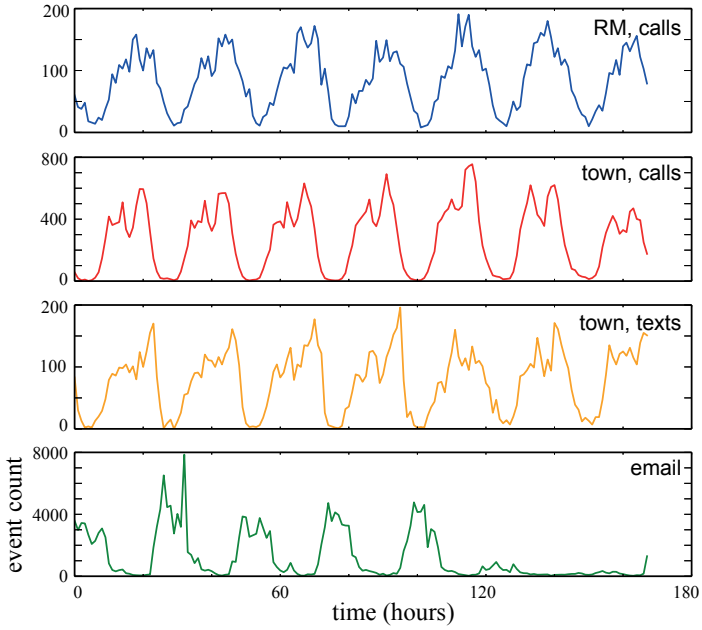

Figure 3.1. Number of events per hour for each day of week over a period of 8 weeks in 4 datasets. From top to bottom: calls in Reality Mining, calls and text messages in a small European town, and emails. In all cases we can see cycles of about 24 hours in the event counts. For the small town, there are differences in activity levels for calls and and text messages. The email dataset has less activity during weekends. Reprinted from [128].

at these results, we have to keep in mind that these rhythms are a superposition of many individual rhythms. However, we should avoid the ecological fallacy and not assume that the system's aggregate rhythm will be similar to the individuals' rhythms. Different people can have very different rhythms and still produce what we see as the average rhythm. To see how much individuals differ in their weekly communication rhythms, we next look at rhythms built in a similar fashion, except that we will aggregate each individual's events separately. To build the individuals' daily communication patterns or *daily rhythms*, we follow the same procedure, however this time we aggregate all events of one individual within the whole study period into 24 bins of width one hour. The aggregated counts are normalized to unity. The result can be seen as a probability distribution of communication events at different times of the day.

3.3 Persistent behavioral patterns of individuals

There are persistent individual differences in behavior, interaction, and communication. This indicates that each person has a "strategy" or "agenda" which guides their behavior. In this section, we look at some examples of human behavior which have been shown to persist by means of digital traces.

In [129], Saramäki *et. al.* use the UK students dataset described in Chapter 2 to show that individuals' communication with others follows a certain pattern, the so-called *social signature*, that is rather persistent in time even when there is a high level of turnover in their personal networks. This means that the students have a certain broad pattern which they follow in order to allocate their time amongst their social contacts. Those who are emotionally closest to the individual receive large fractions of communication time. The social signatures are rather robust to external changes such as major life events, like moving to another city or finishing secondary school and starting university, which lead to high rate of turnover in their social networks.

Other works have studied strategies with respect to forming new friendships and the decay of old ones. In [130], Miritello *et al.* report the existence of a *social capacity* for each individual. This means that despite the fact that ties between people are constantly being created and destroyed, for each person these tend to happen with a constant rate. They also label the amount of activity in terms of formation and destruction of ties with others, for each individual as their *social activity*. People vary in their social capacity as well as their social activity. Based on the ratio of the social capacity and social activity, individuals can be categorized into two groups: *social keepers* and *social explores*. While explorers tend to have large activity levels compared to their communication capacity, keepers exhibit small such ratios.

In publications I, II, and III, we look at individuals' patterns of communication. We first try to see whether individuals within the same system exhibit more or less the same (or similar) activity patterns. In Fig. 3.2, daily communication patterns of 12 different individuals in 4 different datasets are depicted. We see that the daily rhythm of each individual exhibits very different patterns from the others, even within the same dataset. Similar results can be seen in Fig. 3.3 another dataset (the students' dataset described in the previous chapter) is used. We see in this

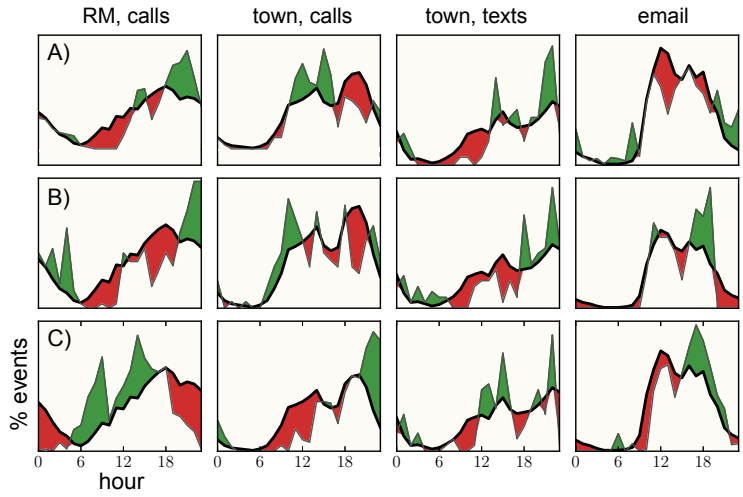

Figure 3.2. Daily patterns of 12 different individuals from four datasets exhibit the diverse nature of individual patterns. The average daily pattern for each dataset is depicted in black. The green/red areas show where individual's pattern is above or below the average. We can see that individual patterns vary significantly from the average pattern of the system. Reprinted from [128].

figure that different individuals differ largely in their daily patterns of two channels of communication: calls and text messages. Some individuals tend to have similar patterns of calls and text messages, while others have different patterns from one channel to another. These results immediately make us think whether these rhythms are only a stochastic manifestation of an individual's activity in a short period of time, or if they are tied to other characteristics of the individual and do not vary significantly in time. To quantify the inter-individual differences in terms of daily rhythms and to examine their persistence for each individual over time, we use a method based on the Jensen-Shannon divergence as we will discuss next.

3.4 Quantification of differences of daily patterns

The Jensen-Shannon divergence (JSD) is a form of Kullback-Leibler divergence (KLD), which measures the distance between two probability distributions. Unlike the KLD, the JSD can handle distributions which have zero-valued elements. JSD for two discrete probability distributions P_1 and P_2 can be written as

$$JSD(P_1, P_2) = H(\frac{1}{2}P_1 + \frac{1}{2}P_2) - \frac{1}{2}[H(P_1) - H(P_2)]. \tag{3.1}$$

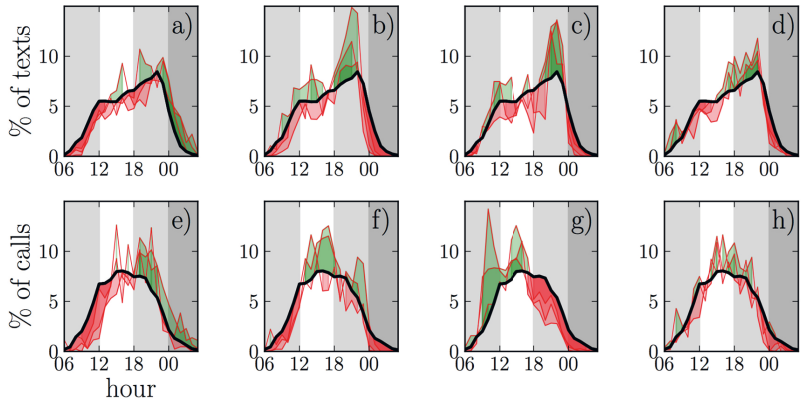

Figure 3.3. Daily patterns of 4 different individuals from UK students dataset. The upper row is based on text messages and the bottom row is based on calls. The average daily pattern for each dataset is shown in black. The green/red areas show where individual's pattern is above or below the average. We can see that each individual's daily rhythm varies from the average pattern. The daily rhythm for one person and one communication channel persists in time. But some individuals have very different daily rhythms for calls and texts. Reprinted from [131].

In our case, P_1 and P_2 are two daily patterns such that $P_i = p_i(t)$ and $p_i(t)$ is the fraction of events in each time interval. H is the Shannon entropy, which is defined as

$$H(P) = -\sum p(t)\log(p(t)). \tag{3.2}$$

Now that we have a measure of the difference between two daily rhythms, we build for each individual a *self* and a *reference distance*. The self distance (for a given individual) is measured by finding the JSD between the daily rhythm of that individual in a given time period with the daily rhythm of the same individual measured in another time period of similar length. This is repeated and averaged over all pairs of consecutive time periods. For example, in publications II and III (UK students dataset) we have divided the full data collection period of 18 months to three time periods of 6 months each. Then, the self distance for individual A is computed between period 1 (months $1-6$) and period 2 (months $7-12$), and the same for periods 2 and 3 (months $13-18$). If we have N consecutive time periods, the self distance for the individual can be written as their average:

$$d^A_{self} = \frac{d_{1,2} + d_{2,3} + ... + d_{N-1,N}}{N-1}. \tag{3.3}$$

To have a reference distance to compare self distances to, we also build a reference distance for each individual, by comparing the distance of the individual's daily rhythm in one time period with the daily rhythms of all

other individuals in the study in the same time period. So if we have M individuals in the study, for each individual in each time period we will have $M-1$ reference distances, and a total of $\frac{M(M-1)}{2}$ reference distances can be calculated within one time period. If the whole study time is divided into N consecutive time periods, we will have a total of $N\frac{M(M-1)}{2}$ reference distances. The total number of self distances will be equal to the number of participants, M. Fig 3.4 depicts the results from publica-

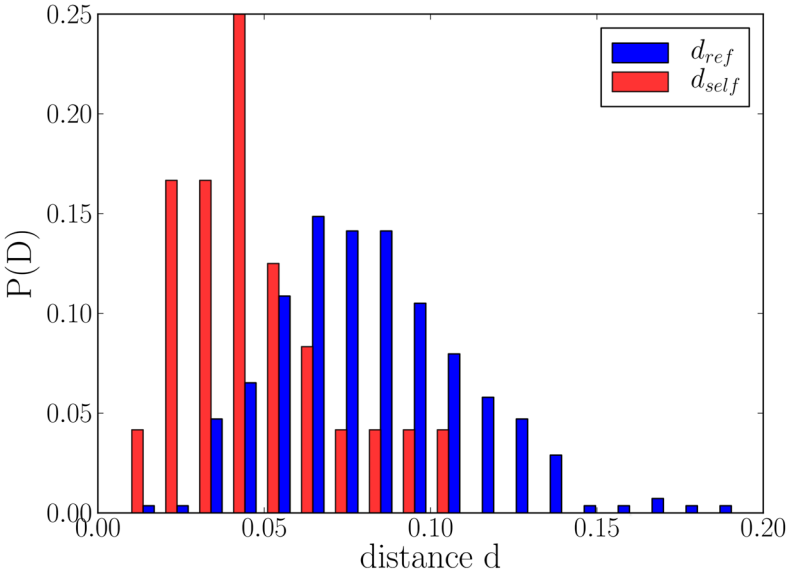

Figure 3.4. These two histograms show the self and reference distances for all individuals (in red and blue respectively). The self distances are clearly smaller than the reference ones. To calculate reference distances, daily rhythm of each individual is always compared to others in the same time period. But, for self distances we compare one's rhythm in different time periods. The smaller self distances in comparison with reference distances on average, shows that changes in the daily rhythms of individuals is rather small over time (persistence of daily rhythms) and that daily rhythms vary among individuals. Reprinted from [132].

tion II, comparing self distances and reference distances for daily rhythms built based on call data from UK students dataset. It is seen that the self distances are smaller than the reference distances, indicating a smaller level if variation between one individual's consecutive intervals. We did a similar analysis for different channels of communication in the same dataset (publication III), as well as for other datasets (publication I), and observed a similar outcome. Based on results from these three publications we come to these main conclusions on the daily rhythms of communication of individuals:

1. Individuals exhibit daily rhythms of communication which are persistent in time.

2. Different individuals have different daily rhythms of communication.

3. These rhythms can be seen in different systems and across multiple channels of communication.

4. Each individual does not necessarily have a similar daily rhythm in different channels of communication, even though the daily rhythm is persistent for each of the channels.

One has to keep in mind that there is turnover in social networks of individuals, meaning that they do not always have the same set of people in their personal networks. Still, the individuals exhibit these persistent behavioral patterns. Even though part of this could possibly be explained by homophily effects, meaning that individuals perhaps tend to substitute previous friendships with new ones which share similarities, part of this persistence is probably rooted in people's internal characteristics. Another observation from our studies of daily rhythms is that the timing of periods of inactivity or low activity is almost as interesting as the times of activity. We see in most cases that there is almost no activity at night, but the position of the start and the end of inactivity varies from one individual to another. One of these persistent features, which is especially important when studying temporal activity patterns, is the individual's so-called "chronotype". In the next section, we explain what chronotypes are and how they are measured. Later on, we discuss findings from publication V regarding identifying chronotypes of individuals based on their digital activity.

3.5 Chronotypes; persistence of sleep and rest time preferences

3.5.1 What are circadian rhythms?

Almost all living cells on Earth follow a close to 24-hour rhythmicity [94, 133]. The word *circadian* comes from Latin words *circa* which means "about" and *dies* meaning "day" [134]. Circadian rhythms are endogenous

and self-sustained biological oscillations which exist even in the absence of light or external clocks [93, 96]. Humans, similar to all other living organisms, exhibit these rhythms at different levels: in their metabolism, physiology, as well as their behavioral patterns [93]. All these oscillations within the body are put in sync by a central pacemaker (master clock) called the suprachiasmatic nuclei (SCN) located in the hypothalamus [135]. SCN regulates the timing of all functions in the body and syncs the body with the outside world using external cues called *zeitgebers* [136]. This process is called *entrainment* [137]. The most important of the zeitgebers is the dark-light cycle [138]. Prior to the industrial age, light mainly came from natural sources, and therefore the daily lives and activities of humans were highly dependent on it. However, artificial lighting gave people the opportunity to be active during a larger part of the day. Murray Melbin, the author of "Night As Frontier" [139], writes: "Time, like space, is part of the ecological niche occupied by a species. Although every type exists throughout the 24-hour cycle, to reflect the way a species uses its niche we label it by the timing of its wakeful life." So, in the same way that humans have spread spatially on Earth (and beyond), they have also been able to push the frontiers of their temporal presence with the help of artificial lights. This has resulted in alterations in humans' sleeping patterns, which shows the importance of the external cues in the process of entrainment. In addition to light, another important set of external cues are social zeitgebers, which are consequences of our social lives, and the existence of man-made schedules and designated hours for working, eating or socializing. The digital revolution has also widened the realm of possibilities of activity, such as allowing socializing at all hours. These extra possibilities are not always advantageous to humans [140]; usage of artificial lights late at night, especially from digital devices, causes disturbances to individuals' circadian rhythms, which can in turn enhance the chance of certain diseases such as cancer and obesity, and reduce life expectancy [135, 141].

Circadian rhythms were known and observed since thousands of years ago [95]. The first experiments were on plants in the 18^{th} century, when leaves were observed to have endogenous rhythmic movements. However it took approximately another 200 years for scientists to study circadian rhythms in humans [142]. Since then circadian rhythms remain a topic of interest and various aspects of them have been explored. For example, different studies have shown that the phase of entrainment of circa-

dian rhythms varies significantly among individuals. There are morning-active and evening-active people, and this is not only reflected in their times of sleep and wakefulness, but also in their internal body processes such as body temperature and metabolism.

3.5.2 What are chronotypes?

In the past two decades, there has been extensive research into categorizing individuals based on their circadian typology, and to find out how circadian rhythms relate to other individual characteristics such as age, gender, health, personality traits, and academic performance [143–145]. In these typologies, individuals are divided into groups referred to as *chronotypes* based on their propensity to sleep at different hours within the day-night cycle. This property (also called *morningness-eveningness*), which is a result of phase difference of entrainment for different individuals, is typically measured by means of questionnaires exclusively designed for this purpose. These typologies commonly divide people into three chronotypes. morning-type (MT), intermediate- (IT) or neither-type (NT), and evening-type (ET). Even though the questionnaires often use hard thresholds to separate these types, it is important to note that in reality we observe a continuum of rhythms which follow a normal distribution in the general population, with morning and evening types each comprising around 25% of the distribution [146, 147]. The extremes of these two types vary significantly in the phase of their sleep-wake cycle; one type tends to wake up around the time when the other type is going to bed [147].

One of the earliest questionnaires to identify chronotypes which is still one of the most widely used surveys has 19 multiple-choice questions. It is called the Morningness-Eveningness Questionnaire (MEQ). This questionnaire was designed by Horne and Östberg in 1976 [148], based on an earlier version in Swedish [149]. MEQ has been validated against internal circadian rhythms such as core body temperature [150]. Much effort has also been put into optimizing this questionnaire and adapting and validating it for different cultures and languages [151]. It has also been shown that small number of items in the questionnaire explain most of the variance [144]. In order to make this questionnaire more suitable for large-scale studies, a reduced version of it was designed in 1991 [152]. This reduced version called rMEQ has since been translated to several languages, and adapted to different cultures, with the number of questions ranging between 3 and 6 [151, 153–155]. In addition to cultural and lan-

guage differences, which make it difficult to use the same questionnaire for different populations, issues have been raised regarding the validity of these questionnaires for cohorts such as shift workers. Neither the MEQ nor rMEQ take into account these less common lifestyles which can significantly influence an individual's resting and sleeping habits. Another very well-known measure which has addressed this issue is the Munich Chronotype Questionnaire (MCTQ) [156]. Results obtained with different questionnaires highly correlate with one another, even though they measure different aspects of the chronotype [144]. For example, MEQ measures the individual phase preferences over the 24-hour cycle, whereas MCTQ evaluates the phase of sleep positions [157].

3.5.3 Inferring sleep from digital footprints

In the recent years, two categories of studies have emerged which try to make use of modern technologies and the data produced by them to gain knowledge about individuals' sleep and circadian rhythms. The first group of studies focuses on inferring various sleep parameters, such as duration, sleep and wake times, and sleep quality, by means of passive data collected from devices like mobile phones. These studies use data streams such as accelerometers, ambient light and noise, screen-on and off events, or a combination of the above to determine per-individual sleep parameters for each instance of sleep [158–160]. For example, in a recent study [161], Cuttone *et. al.* use the Copenhagen Networks Study dataset and apply a Bayesian method to screen-on and off events collected from mobile phones to find the probability of being asleep or awake at each point in time for each individual. A second set of studies which are labeled as "circadian computing" use questionnaires and determine the chronotype of study participants first, and then study how a person's chronotype correlates with the way they use their phone, for example app usage [162], their sleep parameters inferred using similar methods as the first study group [163], or unhealthy sleeping patterns [164].

3.5.4 Identifying chronotypes from digital activity rhythms

In publication V we use the Copenhagen Networks Study dataset (see Fig. 2.2), to identify each participant's chronotype. Unlike some of the previous studies, we do not focus on estimation of sleep and wake up times for single days, but our method is based on the assumption from chronobi-

ology literature that even though a person's chronotype may change over the course of their life, these changes are not frequent [165]. So it is safe to assume that in our homogeneous population of young adults, each person's chronotype remains more or less the same within our analysis period of one year. We look at weekly digital activity patterns of individuals and by comparing each person's pattern with the population average, we identify their chronotype. By digital activity, we mean the time each mobile phone screen turns on, which may be considered a better proxy of being awake than times of calls, which do not necessarily happen close to sleep and wake up times. The daily activity patterns of individuals commonly differ in weekdays compared to weekends when they are not constricted by work or study schedules [166]. To identify chronotypes we only use activity patterns from Monday to Thursday to avoid the influence of less strict schedules towards the end of the week. After building each individual's weekly activity pattern, we compare it to the population average in early hours ($5-7$ am) and late hours (midnight-2 am). We categorize 20% of students with highest activity level in the late (early) hours and low activity in the early (late) hours as ET (MT). In Fig. 3.5, we can see the weekly digital activity rhythm of two different individuals (one morning-type and one evening-type), together with the population average rhythm. It is evident that the rhythms of the two types clearly deviate from the population average in the early and late hours.

3.5.5 Chronotypes in social context

In addition to the assessment of individuals' circadian typology, the epidemiology of chronotypes and how they correlate with other personal characteristics have widely been studied [144]. Different chronotypes have been linked to differences in mental and physical health [168–171], academic performance [172, 173] and the level of income [174], among other things. Their distributions have also been studied for different age groups [144, 175], genders [143, 144], and other sociodemographic properties [144, 176]. Despite the importance of circadian typology in many aspects of individuals' lives, their implication in the social life has little been studied. We know that the chronotype of a person can affect their life in health as well as in disease. This naturally makes us wonder how these typologies influence the social life and behavior of individuals.

In article V, we look at homophily and centrality measures within the network of participants in the Copenhagen Networks Study dataset, and

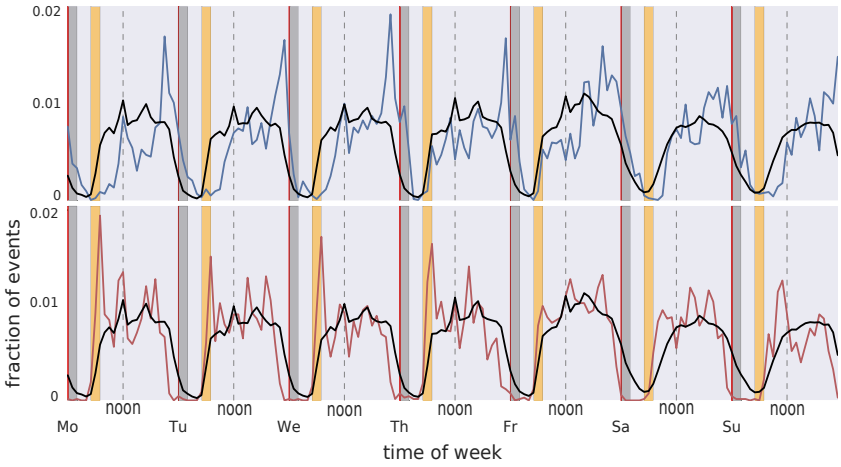

Figure 3.5. Weekly rhythms of one morning-type individual (red), one evening-type (blue), and the population average rhythm (black). Adapted from [167]

show that evening-types tend to communicate more among each other. In contrast, such homophily is absent for morning-types. Using different measures of centrality we show that evening-types take a more central position in the social network of participants (see Fig. 3.6).

In this article, we show that the chronotypes of individuals can have significant implications in the way as well the times they communicate with others. This can have important implications in public health or epidemiology and indicates that future research in this direction is necessary. Also, in the future, the method we have developed for identifying chronotypes, can be used for other large empirical datasets. This can circumvent some of the common issues with questionnaires for identifying chronotypes, because it is based on actual activity of individuals rather than the "typical activity" that they report.

3.6 Communication strategies of individuals

The same way a social system is a superposition of patterns of individuals, one person's behavior can be thought of as a superposition of several features of the individual, such as age, gender, or chronotype. Here, we explore how these individual characteristics affect the temporal patterns

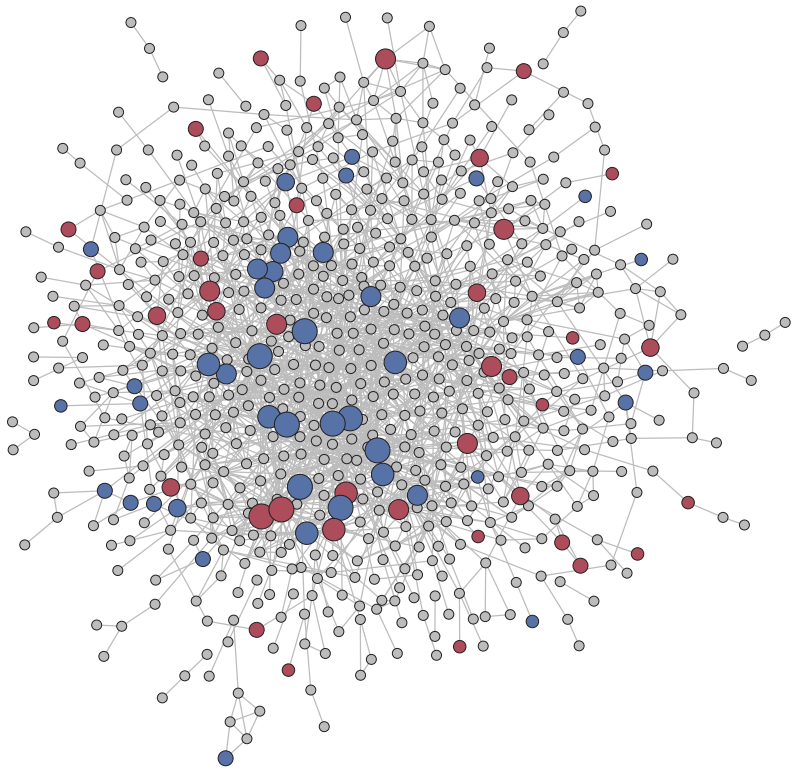

Figure 3.6. Social network of participants in Copenhagen Networks study dataset. Red and blue nodes represent MT and ET individuals respectively. All the other participants are depicted in gray. The size of the blue and red nodes correspond to their core number (more detail in publication V). Reprinted from [167].

of their communications with others. In communication, the nature of the relation between people is also important. For example, people do not usually communicate with friends and family in the same manner [76].

Even before the age of large empirical datasets, there has been interest in understanding how individuals with different characteristics differ in the way they interact with others. For example, differences between communication partners of males and females and their talkativeness has been extensively studied [177, 178]. Similar correlations have been found in large digital datasets. For example, in [179], the talkativeness of different genders as well as the physical proximity during interactions has been studied using data collected from digital proximity sensors. Other works have found links between the ways people interact and communicate with one another and their gender and age [67, 180, 181], emotional closeness [76, 129], and economic status [182]. Such studies are possible

when passive data collected from digital devices is augmented with sociodemographic data gathered from the subjects (such as age, gender or their emotional closeness with the members of their social network), for example by means of questionnaires.

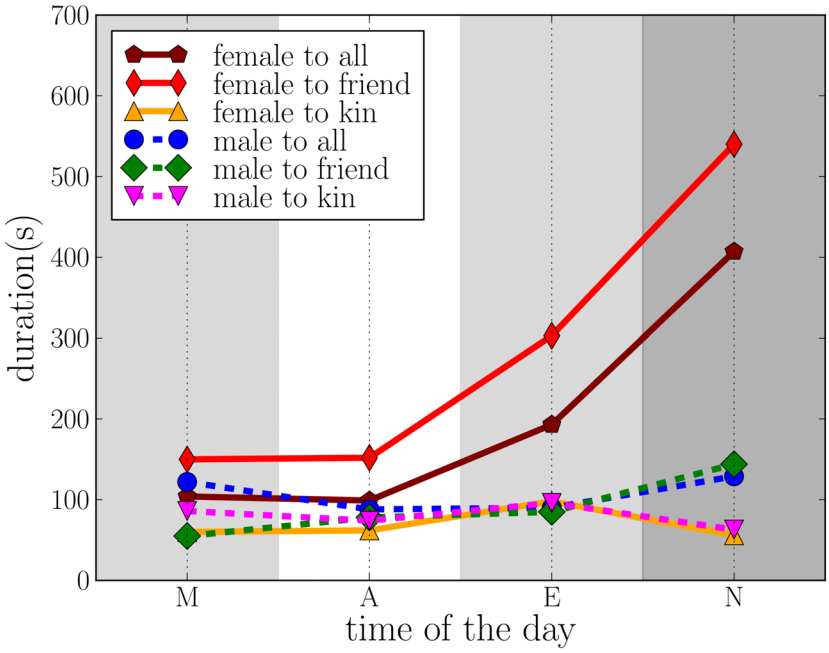

Figure 3.7. Average duration of calls between study participants and their friends and different daily time bins: Morning, Afternoon, Evening, and Night. This is based on the UK students dataset. Reprinted from [132].

As already established, this thesis explores temporal activity patterns of individuals, with a focus in temporal communication patterns. In publications II, III, and V, we have looked at different properties of individuals and their social network, and how these affect communication patterns at different times of the day. In publications II, and III, we look at communication data from different channels (calls and text messages) and see how the amount of communication varies, both in terms of number of events and duration (for calls), in 4 different sections of the day: morning, afternoon, evening, and night. We show that the amount of communication at different hours varies for communications with friends as compared to kin. In Fig. 3.7, we see the average duration of calls made by males and females to their kin and friends. Data from 18 months for participants of the UK students dataset have been used for this plot. We see that duration of calls is on average shorter for communication with kin, at all hours, but especially at night, while calls with friends tend to get longer

at late hours. Also, communication within the same gender and between genders show variations depending on the time of day. For example, calls at night from females to males tend to be significantly longer than calls from males to females or calls with the same gender (see 3.8.)

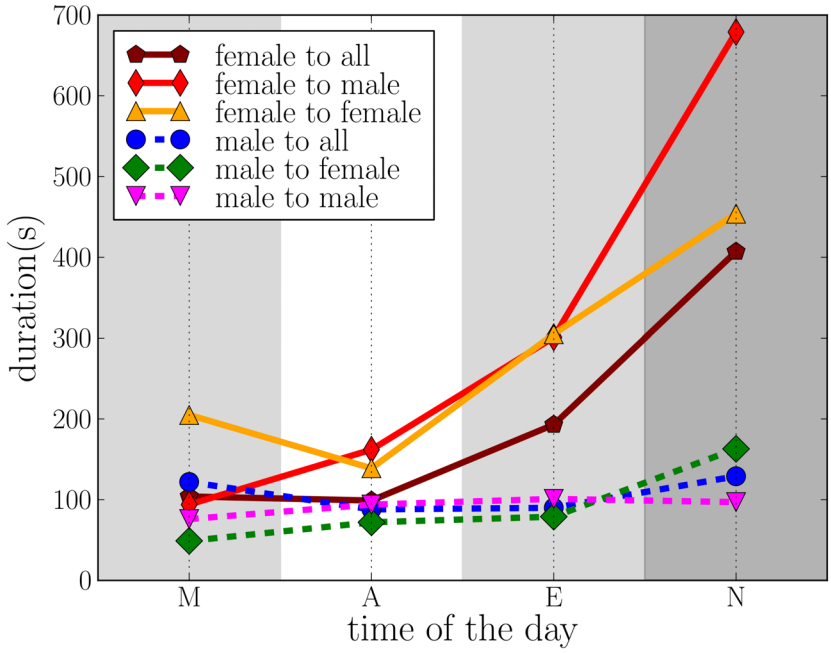

Figure 3.8. Average duration of calls between different genders in four different daily time bins: Morning, Afternoon, Evening, and Night. This is based on UK students dataset. Reprinted from [132].

In article V, we build the social network of each individual in the Copenhagen Networks Study dataset (using their communication data) and show that evening-type individuals maintain larger social networks, but spend less time communicating with those in their network. In Fig. 3.9 we show how the communication activity for each chronotype is spread throughout different times of the day, both in terms of number of calls and duration of calls.

It is known that values of different network centrality measures correlate with node degrees. As a result, it is not surprising that different centrality measures for ET and MT individuals vary since they have different personal network sizes. But the homophily between ET individuals and lack of it for MT ones cannot be explained by centralities or personal network sizes. Since chronotype of individuals affects the times they are active, this can have important implications when studying temporal networks, for example in studying spreading process on networks. This is an

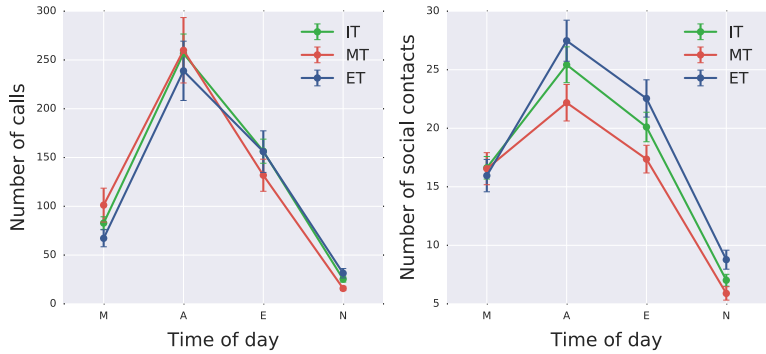

Figure 3.9. Left panel: number of calls at different times of day (Morning, Afternoon, Evening, and Night), between MT, ET, and IT individuals and their social contacts. Right panel: number of people that MT, ET, and IT individuals are in contact with (through phone calls) at different times of day (Morning, Afternoon, Evening, and Night).

area which definitely merits further investigations in the future.

4. High-resolution data collection studies

In Chapter 2, we discussed data collection experiments that rely on modern technologies, specifically mobile phones. The aim of such studies is to add to the depth of the research data compared to what comes from sources such as social networking websites or wikis, whose data are not collected with some particular hypothesis in mind. For these data provided by third parties, researchers often do not have detailed information about users, such as their age, gender, personality, or health status. Also, one has to note that for most of the data sources which are not open data, researcher access is subject to severe scrutiny from the company or the organization which owns the dataset. Running data collection experiments can solve these issues to some extent. Since participation in such experiments is voluntary, participants sign consent forms prior to the study so that researchers can have a right to use their data. Furthermore, the design of the experiment is typically scrutinized by an ethics committee, whose approval is required for conducting the experiment. Studies such as the Copenhagen Networks Study or the Reality Mining study (both described in Chapter 2) are good examples of this type of experiments. In the recent years, there have been many other data collection experiments. Some new studies use devices other than mobile phones to collect data with very high-resolution for a particular goal. A good example is the "SocioPatterns" project which is discussed in more detail in the next section. In addition to that, some other projects are dedicated to building data collection apps or platforms which can be used by other research groups (*e.g.* the AWARE framework or the Beiwe platform, both discussed in the next section). The increasing number of data collection studies makes it important to address issues related to the data and study participants such as privacy, ethics, and consent, as well as more scientific and technical issues such as reproducibility of results, combining data sources, data

analysis techniques, etc.

In publication VI we report on design features which we have used for our prototype data collection platform. We believe they can aid future platforms to have better features. In most data collection studies, publications are focused on the results of the study rather than on the process of designing and building the data collection platform and other technical issues. For making research in this field more efficient, so that not every new data collection study has to start from scratch, this needs to change and more studies should report on their procedures.

In this chapter, we first discuss some recent experiments and data collection platforms, then tackle broader subjects and talk about important issues in designing data collection platforms through the example of Nima, the "Non-Intrusive Individual Monitoring Architecture", a platform designed and developed at the Complex Systems research group at Aalto University, Finland.

4.1 Existing projects

SocioPatterns

SocioPatterns [183] is not just one study, but rather a research initiative which was formed in 2008. The studies of this initiative utilize wireless devices embedded in wearable badges, which collect high-resolution data on proximity of individuals. These radio-frequency identification (RFID) sensors are low-cost devices which recognize other similar devices in their proximity by exchanging radio packets [184]. Human face-to-face interactions play a major role in, among other things, contagion and spreading processes [185, 186]. Collecting data on these interactions can be done using several methods, such as contact diaries, mobile phone Bluetooth sensors, or observational studies [187]. However, these methods are prone to different errors (*e.g.* recall biases for diaries) or lack precision [188]. RFID sensors (as well as other proximity sensors) have their own shortcomings: for example they need to be used in closed settings with high rate of participation so that most interactions are captured [187]. However, by using custom-made RFID sensors with good calibration and fine-tuning of the sensitive range, one can achieve high interaction capture rates [187]. The SocioPatterns project has studied contact patterns and

their implications for spreading processes in various closed settings such as conferences [189], hospital wards [190], museum exhibitions [191], and schools [192, 193].

Beiwe

Beiwe is a data collection platform designed in the Onnela lab at Harvard T.H. Chan School of Public Health. This framework offers apps both for Android and iOS devices and it is designed for data collection for *digital phenotyping* studies. Digital phenotyping is a term coined by the same group as "moment-by-moment quantification of the individual-level human phenotype *in situ* using data from personal digital devices" [194]. Beiwe is meant to be used by different research groups working in the behavioral sciences and the public health domain, for example in monitoring patients after surgeries, or studying patients with mental disorder [195, 196]. To use the Beiwe platform for studies, the Onnela lab envisions three models [197]: 1) direct collaboration with the lab, 2) using the app as a service subject to a fee, 3) using an open version of the platform. At the moment, the second and the third method have not yet been made available [197].

AWARE

AWARE, or the Aware Framework [198, 199], is a set of tools for running studies using primarily mobile phones. It was originally created by Denzil Ferreira at the University of Oulu. It consists of an Android app (Oulu, Finland), an Apple iOS app (developed independently at Keio University, Japan), and a server to manage studies, collect data over the Internet, and allow researchers to access data. Studies that have used AWARE have mainly been conducted from the point of view of Ubiqutous Computing research [200, 201]. All components of AWARE are open source, and have been adapted and improved as part of the research at Aalto University. AWARE is not designed for any specific research question, but rather as a tool which can be used by various types of users. It has been designed with three types of users in mind: individuals recording their own data, scientists using the framework to carry out experiments, and app developers who want to use AWARE as part of their projects [202].

4.2 The ethical framework of data collection

Modern data collection methods collect rich data on humans, but at the same time, with the data comes responsibility. Data should be collected, transfered, stored, and accessed in a way that the privacy of participants is not compromised. New technologies and new data collection methods call for revisiting the ethical framework of data-based research. However, the technology is moving forward so fast that the theory behind the ethics of big data is still lagging behind [203]. There is a wide range of issues which have to be discussed and studied in this domain. Here we focus on a few topics which are relevant to our data collection work.

As mentioned in the first chapter, the advent of widespread computation and data analysis has irreversibly changed society. In particular, the relationship between people and data as well as the general view of what is considered as private has changed dramatically. Privacy is a principle which states that individuals should have the right to express how information about themselves can be shared, or "the right to be left alone" [204–207]. Clearly, this concept is intrinsically connected to the types of ways which information can be shared. The modern data-centric world allows data to be shared extremely easily, as a consequence our relationship with privacy has changed [208]. These days, it is routine for individuals to give data on their whole lives to large social media sites for the convenience of sharing and socializing. However, information is given under very broad terms which allow the sites to use the data in almost any way, including directly advertising to, selling, and affecting the life of the person. While people still value privacy, this shows that many will allow wide use of their data in return for useful services—even if those using the data do not have the individual's best interests in mind.

Science has a much stricter ethical and legal framework. Basic ethics of human experimentation, such as the Declaration of Helsinki, assert that all research subjects should have certain rights when undergoing human experiments [209, 210]. In particular subjects have a right to informed consent to any procedures done on them, which is often interpreted as the right to control secondary use of their data. Combined with the ethical requirement to have all projects pre-approved, this would seem to eliminate, or make difficult, the possibility of collecting data and reusing it for follow-up analysis later. However, much of the ethical framework of human experimentation was created with a focus on medical experi-

mentation which directly affects the human body. Pure data collection has no risk to bodily integrity, and passive data collection has a much lower amount of personal risk since there is, in fact, no experiment being formed: the data which is collected is simply that which could be collected for other purposes. The chief risk of these types of experiments is for the privacy of subjects, should data be handled insecurely and spread publicly [211]. This section should not be taken to imply any lack of ethical issues while doing data collection studies. Instead, they are of a fundamentally different nature because the risk occurs after the data is collected, not directly to the subject during the project.

However, recent legal changes provide some benefits for science. In the European Union, the General Data Protection Regulation (GDPR) will soon come into effect [212]. This provides explicit acknowledgement that consent may be given for scientific purposes within broad fields of science, which is intended to include follow-up research. Also, the law makes explicit that data, if anonymized, is no longer considered personal and can be used without limitation of the GDPR [213], though this does not eliminate the need for other ethical considerations. Still, it is only natural that those who approve human experiments are conservative in their approvals. Good experimental design and tools, such as those we discuss in the next section, will allow the optimal use of data.

In addition to general considerations, there are considerations which relate to the specific uses and forms of data which is being collected. Not all data is equal: some forms can have more ethical risks, primarily privacy.

Previous studies have shown that if only direct identifiers are removed from data, individuals can still be identified by means of reverse engineering [214, 215]. Various studies have shown that individuals tend have unique ways of moving, using search engines, or using their credit cards. One way of addressing this issue is to aggregate the data. Aggregation can mean aggregating data from several people, or for example binning the timestamps of events of a single person into large time intervals. It should be noted that anonymization is a very complex problem, and should be carefully tackled on a case-by-case basis [216].

4.3 Niima: a digital data collection platform

Niima is a digital data collection platform which can be used for any type of study which collects individual-level digital data from different sources.

The core of Niima is the Koota server [217] which collects data and contains a dashboard for managing data sources and different experiments. The server can be coupled with different devices and apps. It is designed to be very agnostic to different data sources, so that it is easy to add and remove data sources at any time, including multiple identical devices for participants. To run an experiment using Niima, it is enough that researchers create a new study on the server, assign each participant with an ID, then assign devices to participants. The server collects all the data for each device together and stores it by device ID. Each device is linked to a participant (more details in publication VI) through a flexible and privacy-preserving access control layer. This ensures that devices and participants cannot be easily linked together and therefore helps to preserve privacy. Studies can also be run so that participants manage their own devices and data, meaning that researchers never have the ability to directly link data to participants. While this is not suitable for all projects, it can greatly increase the overall level of privacy protection afforded to subjects.

Niima was originally designed to be used in a project course at Aalto University. However, it was later expanded to be utilized in larger settings, such as data collection from the general population or in clinical studies. At the moment Niima is used in two separate studies, at the Helsinki University Hospital (HUH) and in a neuroscience study at the Department of Neuroscience and Biomedical Engineering at Aalto University. One of the use cases of Niima, mental health studies, is explained in more detail at the end of this chapter. From the beginning, Niima was designed in a way that it can satisfy three primary principles: built-in privacy features, flexible data sources, and fine-grained access control which assists in data minimization. Next, we explain these features in more detail.

4.3.1 Privacy by design

If a system is not designed from the beginning with privacy in mind, it can be very hard to add it later [218]. This is because privacy requires minimizing and compartmentalizing information, and if a system is designed in the most straightforward way this will not be the case. The first step of this is the user vs. device vs. data system described above. With this system, we are assured that data is not essentially tied to a single user. When data is imported, any direct identifiers are stripped.

Data undergoes more processing when it leaves the server: this limits the data which can be provided to any one study, so some studies may receive less data than others. Furthermore, this processing can apply arbitrary transformations when the data is extracted so that, for example, extra anonymization can be applied.

4.3.2 Flexibility of data sources

It is often necessary to be able to collect different types of data from various devices, to have passive as well as active data (which requires user's engagement). To make this type of flexibility possible, a platform must be designed so that it does not presuppose any particular data model or interface. The Koota server is capable of receiving arbitrary data and simply storing it for later processing. Data is only processed when it is extracted. This also allows data to be collected even without fully understanding it, so that the relevant information can be extracted later. This is especially important when integrating third-party devices. Koota has been integrated with three Android applications, two Apple iOS applications, standalone fitness trackers, social media sites, and online surveys. By collecting all data in one place with a common framework, researchers do not have to deal with different databases separately and go through the cumbersome task of overlaying the datasets, which is not only challenging but also compromises the privacy of study participants.

4.3.3 Flexibility of access control

In the past when researchers worked with datasets which were relatively small, they could manage, curate and archive the data themselves, but with growing number of data sources and large amount of data, these tasks are becoming extremely challenging and and can hardly be longer handled by researchers themselves and need their own professionals [15]. This is especially true with modern personal data-driven research, where it should be the goal of researchers to not deal with any personal data at all so that data handling can be minimized. Furthermore, as anybody who has ever engaged in data collection in any form knows, data handling can get very complicated and messy very quickly. Because the data is legally and ethically protected, this must be minimized. Niima makes a clear division of different roles. Full access to data is limited to administrators, and even they do not normally interact with the full data. Per-user, per-

study, and per-data type data access is only given to certain researchers, and all data is processed through converters to apply extra filters for additional privacy. Studies may either be blind, where the managers do not know identities of participants, or certain researchers may be given access to manage the devices and identities of subjects. These two roles can even be separated, so that no person can both know the identity of a subject and access their data. With fine-grained controls such as these, it can be much easier to satisfy legal and ethical rules by default.

4.4 Data collection for mental health studies

One specific use case of Niima is mental health studies. Mental health illnesses, which are one of the major causes of disability worldwide [219], are very challenging to diagnose, control and treat. One reason for this is the fact that mental disorders do not have clear biomarkers similar to physical problems [220]. The most common approach in psychiatry for diagnosis and further follow up of patients is visiting a professional. Psychiatrists use structured and semi-structured interviews with patients (and at times also their relatives) to identify and assess the problem [221]. This method suffers from similar issues as surveys as a data collection method in social sciences: it completely relies on people's accounts and the answers can be subjective. Also, it is prone to memory biases, especially for patients with mental health disorders, since the very source of the problem adds further concerns regarding the reliability of autobiographical accounts. The second problem, which exists also for survey data, is that data points are rather limited because data collection is extremely labor-intensive. The field of computational psychiatry, which is still in a nascent stage similarly to computational social science, takes advantage of new methods in collecting rich behavioral data from individuals by means of wearable devices such as mobile phones. This method, even though still facing many challenges in practice, has several benefits: 1) Passive data, which does not require any interaction with the user, can be collected from patients continuously (digital phenotyping). 2) Data collection can be expanded to other devices for measuring variables whose benefit is already known to psychiatrists, for example actigraphy or using ballistocardiographic bed sensors to measure sleep. 3) Passive behavioral and physiological data collection from multiple devices can be augmented by data that needs active engagement of the patient. For example, Ecological Mo-

mentary Assessment (EMA) is a method which has been used for decades now (both in non-digital and digital form) to collect data on the mood and the state of the patient with high frequency (multiple times a day) to provide psychiatrists with more fine-grained data [222]. This method can be combined and integrated with passive data collection. The outcome of this can directly be used by psychiatrists, but it can also serve as ground truth data for data scientists who try to find meaningful biomarkers from passive data. 4) In [223], Jain *et. al.* introduce the concept of "digital phenotypes" (which should not be confused with digital phenotyping). Digital phenotypes are phenotypes which are a result of the existence of digital devices. For example (unusual) activity on social media is an example of such a phenotype. By linking the data from different social media, data collection platforms can collect data on digital phenotypes as well. Collecting digital data (both passive data as well EMA data) is becoming a more common in psychiatric research. In the recent years, many studies have tried to collect data from patients with different types of disorders such as schizophrenia [224], bipolar disorder [225], and major depressive disorder [226]. Preliminary results in this area are promising, and these studies are bound to see a rise in the number in the coming years.

4.5 Future prospects

In recent years there has been an increasing trend in the number of data collection studies. With large amounts of data produced in these experiments, various privacy and ethical issues are becoming more and more sensitive and wider range of data sources available everyday, the design features of these platforms are becoming more critical. In designing data collection platforms, common approaches are to either outsource the design or to make the platform in-house by researchers who are going to use the data. None of these two approaches can guarantee a smooth functioning and re-usability of the platforms. We suggest that the common approach in CSS studies, of experts from different disciplines working together, should be applied for data collection experiments as well. Experts in designing and maintaining digital platforms should become a part of the loop. This way, there can be constant feedback between researchers, users, designers, and administrators. Also, researchers do not need to handle issues such as storing and pre-processing complex data, or to make sure that the privacy of participants is not compromised. In summary, as

data becomes more complex in the future, the science that collects and uses data needs to become more inter-disciplinary.

5. Conclusions

With the advent of fully digital and programmable computers in the mid 20^{th} century, and their ubiquity by the end of the century, human societies have been fully transformed. We now live in connected societies where data and information are the focal points of our lives. Data from digital devices, produced as the result of interaction of individuals with them, serve as digital footprints which contain rich behavioral information on people. This type of data, often large in velocity, volume and variety, are referred to as big data.

The new emerging field of computational social science is a fully multidisciplinary effort to make sense of big data on humans to address questions within different disciplines of social sciences. In CSS, expertise of scientists from the natural sciences is combined with insights of social scientists to harness the big data to solve the societal challenges such as spread of diseases, burden of mental health problems, and cyber security threats. In this thesis, different types of digital auto-recorded behavioral data were used to explore temporal patterns in human behavior such as mobility, communication, and sleeping and resting times. We also introduced a new generation of data collection experiments which use modern devices and tools to collect data in an unprecedented level of detail about humans. Data collection experiments with custom data will increase the pace of research and benefits to society in the future. Today, it is established that data is useful. There is no doubt that large amounts of digital data which are constantly being produced on people can help us study their behavioral patterns. We also know that it is possible to carry out data collection experiments and tailor the study for our research needs and goals. But with big data comes big responsibility. It is our duty as scientists to establish ethical practices of data handling and analysis. Never before in human history has there been so much detail registered on the

lives of individuals. We have to treat data with vigilance and make sure that privacy of individuals is not compromised.

A decade ago, using CDR data could give us a rather complete picture of an individual's communication. Nowadays, with the rise of different messaging apps, online platforms, social networking websites, etc., having access to calls and text messages of a person through the phone operator is no longer a nearly complete subset of the person's interactions. This is especially true for younger generations. Thus, in the future data collection studies will be more and more important for the purpose of research on human behavioral patterns.

There is an increasing number of data collection platforms being built everyday. These platforms are meant to be used by different research groups. This can help to run studies in the future which have very specific goals, with little preparation and minimal setup efforts.

Having more data collection experiments also means that in many cases there will be data coming from many channels and sensors. This calls for constant development of new methods for extracting useful information from data. Common methods in CSS such as network science, even though still useful in many cases, cannot always totally grasp the complexity of such data. Therefore, for uncovering behavioral patterns of individuals, methods such as predictive modeling and deep learning are possible candidates.

We are now at the dawn of a new era which is centered around information. We will undoubtedly see many developments in collection and analysis of data from humans in the near future. There will be need for different fields to work ever more closely with one another and for training a new generation of scientists who are well equipped to work across disciplines.

Bibliography

[1] T. Hughes, *Networks of Power: Electrification in Western Society, 1880-1930*. ACLS Humanities E-Book, Johns Hopkins University Press, 1993.

[2] J. Mokyr, "The second industrial revolution, 1870-1914," *Storia dell'economia Mondiale*, pp. 219–45, 1998.

[3] T. Forester, *The Information Technology Revolution*. MIT Press, 1985.

[4] H. Capron, *Computers: Tools for an Information Age*. Prentice Hall, 2000.

[5] M. Campbell-Kelly, W. Aspray, D. P. Snowman, S. R. McKay, W. Christian, *et al.*, "Computer a history of the information machine," *Computers in Physics*, vol. 11, no. 3, pp. 256–257, 1997.

[6] R. Conte, N. Gilbert, G. Bonelli, C. Cioffi-Revilla, G. Deffuant, J. Kertesz, V. Loreto, S. Moat, J.-P. Nadal, A. Sanchez, *et al.*, "Manifesto of computational social science," *European Physical Journal-Special Topics*, vol. 214, pp. p–325, 2012.

[7] G. Ifrah, E. F. Harding, D. Bellos, S. Wood, *et al.*, *The universal history of computing: From the abacus to quantum computing*. John Wiley & Sons, Inc., 2000.

[8] M. R. Williams, *A history of computing technology*. IEEE Computer Society Press, 1997.

[9] D. Von Jezierski, *Slide rules: A journey through three centuries*. Astragal Press, 2000.

[10] W. F. Brinkman, D. E. Haggan, and W. W. Troutman, "A history of the invention of the transistor and where it will lead us," *IEEE Journal of Solid-State Circuits*, vol. 32, no. 12, pp. 1858–1865, 1997.

[11] C. Freeman and F. Louçã, *As Time Goes By: From the Industrial Revolutions to the Information Revolution*. OUP Oxford, 2001.

[12] V. D. Blondel, A. Decuyper, and G. Krings, "A survey of results on mobile phone datasets analysis," *EPJ Data Science*, vol. 4, no. 1, p. 10, 2015.

[13] J. Grimmer, "We are all social scientists now: how big data, machine learning, and causal inference work together," *PS: Political Science & Politics*, vol. 48, no. 1, pp. 80–83, 2015.

[14] J.-P. Onnela and S. L. Rauch, "Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health.," *Neuropsychopharmacology: official publication of the American College of Neuropsychopharmacology*, vol. 41, no. 7, pp. 1691–1696, 2016.

[15] G. Bell, T. Hey, and A. Szalay, "Beyond the data deluge," *Science*, vol. 323, no. 5919, pp. 1297–1298, 2009.

[16] S. Sagiroglu and D. Sinanc, "Big data: A review," in *Collaboration Technologies and Systems (CTS), 2013 International Conference on*, pp. 42–47, IEEE, 2013.

[17] S. LaValle, E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz, "Big data, analytics and the path from insights to value," *MIT sloan management review*, vol. 52, no. 2, p. 21, 2011.

[18] W. Hoschek, J. Jaen-Martinez, A. Samar, H. Stockinger, and K. Stockinger, "Data management in an international data grid project," *Grid Computing—GRID 2000*, pp. 333–361, 2000.

[19] T. Berners-Lee, D. Dimitroyannis, A. J. Mallinckrodt, S. McKay, *et al.*, "World wide web," *Computers in Physics*, vol. 8, no. 3, pp. 298–299, 1994.

[20] T. Berners-Lee, M. Fischetti, and M. L. Foreword By-Dertouzos, *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor*. HarperInformation, 2000.

[21] M. Castells, *The Rise of the Network Society: The Information Age: Economy, Society, and Culture*. No. v. 1 in Information Age Series, Wiley, 2011.

[22] M. Satyanarayanan, "Pervasive computing: Vision and challenges," *IEEE Personal communications*, vol. 8, no. 4, pp. 10–17, 2001.

[23] "India sure loves whatsapp: 14 billion messages sent on new year's eve." http://mashable.com/2017/01/06/whatsapp-messages-volume-stat-india. Accessed: 2017-08-10.

[24] "The #rio2016 twitter data recap." https://blog.twitter.com/official/en_us/a/2016/the-rio2016-twitter-data-recap.html. Accessed: 2017-08-10.

[25] L. Gualtieri, S. Rosenbluth, and J. Phillips, "Can a free wearable activity tracker change behavior? the impact of trackers on adults in a physician-led wellness group," *JMIR research protocols*, vol. 5, no. 4, 2016.

[26] L. Piwek, D. A. Ellis, S. Andrews, and A. Joinson, "The rise of consumer health wearables: promises and barriers," *PLoS Medicine*, vol. 13, no. 2, p. e1001953, 2016.

[27] O. Banos, C. Villalonga, M. Damas, P. Gloesekoetter, H. Pomares, and I. Rojas, "Physiodroid: Combining wearable health sensors and mobile devices for a ubiquitous, continuous, and personal monitoring," *The Scientific World Journal*, vol. 2014, 2014.

[28] A. Katal, M. Wazid, and R. Goudar, "Big data: issues, challenges, tools and good practices," in *Contemporary Computing (IC3), 2013 Sixth International Conference on*, pp. 404–409, IEEE, 2013.

[29] S. Ghemawat, H. Gobioff, and S.-T. Leung, "The google file system," in *ACM SIGOPS operating systems review*, vol. 37, pp. 29–43, ACM, 2003.

[30] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in *Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on*, pp. 1–10, IEEE, 2010.

[31] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[32] M. Bhandarkar, "Mapreduce programming with apache hadoop," in *Parallel & Distributed Processing (IPDPS), 2010 IEEE International Symposium on*, pp. 1–1, IEEE, 2010.

[33] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster computing with working sets.," *HotCloud*, vol. 10, no. 10-10, p. 95, 2010.

[34] M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M. J. Franklin, *et al.*, "Apache spark: A unified engine for big data processing," *Communications of the ACM*, vol. 59, no. 11, pp. 56–65, 2016.

[35] K. M. M. Thein, "Apache kafka: Next generation distributed messaging system," *International Journal of Scientific Engineering and Technology Research*, vol. 3, no. 47, pp. 9478–9483, 2014.

[36] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan, "The rise of "big data" on cloud computing: Review and open research issues," *Information Systems*, vol. 47, pp. 98–115, 2015.

[37] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine*, vol. 17, no. 3, p. 37, 1996.

[38] T. Mitchell, *Machine Learning*. McGraw-Hill International Editions, McGraw-Hill, 1997.

[39] S. Emmott and S. Rison, "Towards 2020 science," *Science in Parliament*, vol. 65, no. 4, pp. 31–33, 2008.

[40] D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, *et al.*, "Life in the network: the coming age of computational social science," *Science (New York, NY)*, vol. 323, no. 5915, p. 721, 2009.

[41] S. Takahashi, D. Sallach, and J. Rouchier, *Advancing Social Simulation: The First World Congress*. Agent-Based Social Systems, Springer Japan, 2008.

[42] A. Bhattacherjee, "Social science research: Principles, methods, and practices,", 2012.

[43] F. Fowler, *Survey Research Methods*. Applied Social Research Methods, SAGE Publications, 2013.

[44] G. Miritello, *Temporal Patterns of Communication in Social Networks*. Springer Theses, Springer International Publishing, 2013.

[45] V. Colizza, A. Barrat, M. Barthélemy, and A. Vespignani, "The role of the airline transportation network in the prediction and predictability of global epidemics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 7, pp. 2015–2020, 2006.

[46] V. Colizza, A. Barrat, M. Barthelemy, A.-J. Valleron, and A. Vespignani, "Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions," *PLoS medicine*, vol. 4, no. 1, p. e13, 2007.

[47] R. Gross and A. Acquisti, "Information revelation and privacy in online social networks," in *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, pp. 71–80, ACM, 2005.

[48] Z. Lu, X. Lu, W. Wang, and C. Wang, "Review and evaluation of security threats on the communication networks in the smart grid," in *Military Communications Conference, 2010-MILCOM 2010*, pp. 1830–1835, IEEE, 2010.

[49] R. Guimerà, A. Díaz-Guilera, F. Vega-Redondo, A. Cabrales, and A. Arenas, "Optimal network topologies for local search with congestion," *Physical review letters*, vol. 89, no. 24, p. 248701, 2002.

[50] D. J. Ashton, T. C. Jarrett, and N. F. Johnson, "Effect of congestion costs on shortest paths through complex networks," *Physical review letters*, vol. 94, no. 5, p. 058701, 2005.

[51] "Facebook." https://www.facebook.com.

[52] "Google." https://www.google.com.

[53] H. Abdi, D. Valentin, and B. Edelman, *Neural Networks*. No. no. 124 in Neural Networks, SAGE Publications, 1999.

[54] G. Cowan, D. Pines, D. Meltzer, and E. *, *Complexity: Metaphors, Models, And Reality*. Avalon Publishing, 1994.

[55] "Status update: Facebook has 2 billion users. can it reach 3 billion?." https://www.usatoday.com/story/tech/news/2017/06/27/status-update-facebook-has-2-billion-users-can-reach-3-billion/103104200/?siteID=hL3Qp0zRBOc-jBhe3mm4DPVWbtiJ6AmyMw. Accessed: 2017-07-15.

[56] "Twitter." https://www.twitter.com.

[57] "Instagram." https://www.instagram.com.

[58] "Linkedin." https://www.linkedin.com.

[59] A. D. Kramer, J. E. Guillory, and J. T. Hancock, "Experimental evidence of massive-scale emotional contagion through social networks," *Proceedings of the National Academy of Sciences*, vol. 111, no. 24, pp. 8788–8790, 2014.

[60] I. M. Verma, "Editorial expression of concern: Experimental evidence of massivescale emotional contagion through social networks," *Proceedings of the National Academy of Sciences*, vol. 111, no. 29, p. 10779, 2014.

[61] "Wikipedia." https://www.wikipedia.org.

[62] T. Yasseri, R. Sumi, and J. Kertész, "Circadian patterns of wikipedia editorial activity: A demographic analysis," *PloS one*, vol. 7, no. 1, p. e30091, 2012.

[63] J. Török, G. Iñiguez, T. Yasseri, M. San Miguel, K. Kaski, and J. Kertész, "Opinions, conflicts, and consensus: Modeling social dynamics in a collaborative environment," *Physical Review Letters*, vol. 110, no. 8, p. 088701, 2013.

[64] T. Yasseri and J. Bright, "Wikipedia traffic data and electoral prediction: towards theoretically informed models," *EPJ Data Science*, vol. 5, no. 1, pp. 1–15, 2016.

[65] M. Mestyán, T. Yasseri, and J. Kertész, "Early prediction of movie box office success based on wikipedia activity big data," *PloS one*, vol. 8, no. 8, p. e71226, 2013.

[66] G. Krings, M. Karsai, S. Bernhardsson, V. D. Blondel, and J. Saramäki, "Effects of time window size and placement on the structure of an aggregated communication network," *EPJ Data Science*, vol. 1, no. 1, p. 4, 2012.

[67] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, "Structure and tie strengths in mobile communication networks," *Proceedings of the national academy of sciences*, vol. 104, no. 18, pp. 7332–7336, 2007.

[68] F. Peruani and L. Tabourier, "Directedness of information flow in mobile phone communication networks," *PloS one*, vol. 6, no. 12, p. e28860, 2011.

[69] M. Karsai, M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, A.-L. Barabási, and J. Saramäki, "Small but slow world: How network topology and burstiness slow down spreading," *Physical Review E*, vol. 83, no. 2, p. 025102, 2011.

[70] C. Song, T. Koren, P. Wang, and A.-L. Barabási, "Modelling the scaling properties of human mobility," *Nature Physics*, vol. 6, pp. 818–823, 2010.

[71] B. C. Csáji, A. Browet, V. A. Traag, J.-C. Delvenne, E. Huens, P. Van Dooren, Z. Smoreda, and V. D. Blondel, "Exploring the mobility of mobile phone users," *Physica A: statistical mechanics and its applications*, vol. 392, no. 6, pp. 1459–1473, 2013.

[72] Y.-A. de Montjoye, Z. Smoreda, R. Trinquart, C. Ziemlicki, and V. D. Blondel, "D4d-senegal: the second mobile phone data for development challenge," *arXiv preprint arXiv:1407.4885*, 2014.

[73] N. Eagle and A. S. Pentland, "Reality mining: sensing complex social systems," *Personal and ubiquitous computing*, vol. 10, no. 4, pp. 255–268, 2006.

[74] M. Raento, A. Oulasvirta, R. Petit, and H. Toivonen, "Contextphone: A prototyping platform for context-aware mobile applications," *IEEE pervasive computing*, vol. 4, no. 2, pp. 51–59, 2005.

[75] "Reality mining dataset." http://realitycommons.media.mit.edu/realitymining.html.

[76] S. G. Roberts and R. I. Dunbar, "Communication in social networks: Effects of kinship, network size, and emotional closeness," *Personal Relationships*, vol. 18, no. 3, pp. 439–452, 2011.

[77] A. Stopczynski, V. Sekara, P. Sapiezynski, A. Cuttone, M. M. Madsen, J. E. Larsen, and S. Lehmann, "Measuring large-scale social networks with high resolution," *PloS one*, vol. 9, no. 4, p. e95978, 2014.

[78] J. Fournier, *Graphs Theory and Applications: With Exercises and Problems*. ISTE, Wiley, 2013.

[79] B. Bollobas, *Modern Graph Theory*. Graduate Texts in Mathematics, Springer New York, 2013.

[80] J. Scott, *Social Network Analysis*. SAGE Publications, 2012.

[81] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *nature*, vol. 393, no. 6684, p. 440, 1998.

[82] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[83] L. Kovanen, *Computational Analysis of Large and Time-dependent Social Networks: Lauri Kovanen*. Aalto University publication series: Doctoral dissertations, Aalto Univ., 2013.

[84] S. Borgatti, M. Everett, and J. Johnson, *Analyzing Social Networks*. SAGE Publications, 2013.

[85] M. Newman, *Networks: An Introduction*. OUP Oxford, 2010.

[86] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[87] J. A. Dunne, R. J. Williams, and N. D. Martinez, "Food-web structure and network theory: the role of connectance and size," *Proceedings of the National Academy of Sciences*, vol. 99, no. 20, pp. 12917–12922, 2002.

[88] H. Jeong, S. Mason, A.-L. Barabási, and Z. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, p. 41, 2001.

[89] G. A. Pagani and M. Aiello, "The power grid as a complex network: a survey," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 11, pp. 2688–2700, 2013.

[90] S. Dorogovtsev and J. Mendes, *Evolution of Networks: From Biological Nets to the Internet and WWW*. OUP Oxford, 2013.

[91] G. Caldarelli and A. Chessa, *Data Science and Complex Networks: Real Case Studies with Python*. Oxford University Press, 2016.

[92] T. Roenneberg, *Internal Time: Chronotypes, Social Jet Lag, and Why You're So Tired*. Harvard University Press, 2012.

[93] R. S. Edgar, E. W. Green, Y. Zhao, G. van Ooijen, M. Olmedo, X. Qin, Y. Xu, M. Pan, U. K. Valekunja, K. A. Feeney, *et al.*, "Peroxiredoxins are conserved markers of circadian rhythms," *Nature*, vol. 485, no. 7399, p. 459, 2012.

[94] S. Panda, J. B. Hogenesch, and S. A. Kay, "Circadian rhythms from flies to human," *Nature*, vol. 417, no. 6886, pp. 329–335, 2002.

[95] R. G. Foster and L. Kreitzman, "The rhythms of life: what your body clock means to you!," *Experimental physiology*, vol. 99, no. 4, pp. 599–606, 2014.

[96] I. Edery, "Circadian rhythms in a nutshell," *Physiological genomics*, vol. 3, no. 2, pp. 59–74, 2000.

[97] J. Wajcman, *Pressed for Time: The Acceleration of Life in Digital Capitalism*. University of Chicago Press, 2014.

[98] A. Zavada, M. C. Gordijn, D. G. Beersma, S. Daan, and T. Roenneberg, "Comparison of the munich chronotype questionnaire with the horne-östberg's morningness-eveningness score," *Chronobiology international*, vol. 22, no. 2, pp. 267–278, 2005.

[99] K. Allan, *Explorations in Classical Sociological Theory: Seeing the Social World*. SAGE Publications, 2012.

[100] J. Šubrt and R. Cassling, "The problem of time from the perspective of the social sciences," *Czech Sociological Review*, pp. 211–224, 2001.

[101] A. Ghosh, D. Monsivais, K. Bhattacharya, and K. Kaski, "Social physics: Understanding human sociality in communication networks," in *Econophysics and Sociophysics: Recent Progress and Future Directions*, pp. 187–200, Springer, 2017.

[102] S. N. Dorogovtsev, J. F. Mendes, and A. N. Samukhin, "Size-dependent degree distribution of a scale-free growing network," *Physical Review E*, vol. 63, no. 6, p. 062101, 2001.

[103] S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3, pp. 75–174, 2010.

[104] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.

[105] M. Starnini, A. Baronchelli, and R. Pastor-Satorras, "Temporal correlations in social multiplex networks," *arXiv preprint arXiv:1606.06626*, 2016.

[106] K. Donnelly, *Adolphe Quetelet, Social Physics and the Average Men of Science, 1796-1874*. University of Pittsburgh Press, 2016.

[107] P. Sen and B. Chakrabarti, *Sociophysics: An Introduction*. OUP Oxford, 2013.

[108] C. Castellano, S. Fortunato, and V. Loreto, "Statistical physics of social dynamics," *Reviews of modern physics*, vol. 81, no. 2, p. 591, 2009.

[109] M. Buchanan, *Forecast: What Physics, Meteorology and the Natural Sciences Can Teach Us about Economics*. Bloomsbury, 2013.

[110] A.-L. Barabasi, "The origin of bursts and heavy tails in human dynamics," *arXiv preprint cond-mat/0505371*, 2005.

[111] J. Saramäki and E. Moro, "From seconds to months: an overview of multi-scale dynamics of mobile telephone calls," *Eur. Phys. J. B*, vol. 88, p. 164, 2015.

[112] L. Kovanen, "Computational analysis of large and time-dependent social networks; suurten ja aikariippuvien sosiaalisten verkostojen laskennallinen analyysi," 2013.

[113] W. C. Wimsatt, "False models as means to truer theories," *Neutral models in biology*, pp. 23–55, 1987.

[114] A. Barabási, *Network Science*. Cambridge University Press, 2016.

[115] P. Holme and J. Saramäki, "Temporal networks," *Physics Reports*, vol. 519, no. 3, pp. 97–125, 2011.

[116] P. Holme and J. Saramäki, "Temporal networks as a modeling framework," in *Temporal networks*, pp. 1–14, Springer, 2013.

[117] J. L. Toole, M. Ulm, M. C. González, and D. Bauer, "Inferring land use from mobile phone activity," in *Proceedings of the ACM SIGKDD international workshop on urban computing*, pp. 1–8, ACM, 2012.

[118] T. Louail, M. Lenormand, O. G. C. Ros, M. Picornell, R. Herranz, E. Frias-Martinez, J. J. Ramasco, and M. Barthelemy, "From mobile phone data to the spatial structure of cities," *Scientific reports*, vol. 4, 2014.

[119] J. Froehlich, J. Neumann, and N. Oliver, "Measuring the pulse of the city through shared bicycle programs," *Proc. of UrbanSense08*, pp. 16–20, 2008.

[120] D. Villatoro, J. Serna, V. Rodríguez, and M. Torrent-Moreno, "The tweetbeat of the city: Microblogging used for discovering behavioural patterns during the mwc2012," in *Citizen in Sensor Networks*, pp. 43–56, Springer, 2013.

[121] M. ten Thij, S. Bhulai, and P. Kampstra, "Circadian patterns in twitter," *Data Analytics*, pp. 12–17, 2014.

[122] T. Yasseri, G. Quattrone, and A. Mashhadi, "Temporal analysis of activity patterns of editors in collaborative mapping project of openstreetmap," in *Proceedings of the 9th International Symposium on Open Collaboration*, p. 13, ACM, 2013.

[123] E. Gabarron, A. Y. Lau, and R. Wynn, "Is there a weekly pattern for health searches on wikipedia and is the pattern unique to health topics?," *Journal of medical Internet research*, vol. 17, no. 12, 2015.

[124] E. Gabarron, A. Lau, and R. Wynn, "Weekly pattern for online information seeking on hiv-a multi-language study.," *Studies in health technology and informatics*, vol. 228, pp. 778–782, 2016.

[125] J.-P. Eckmann, E. Moses, and D. Sergi, "Entropy of dialogues creates coherent structures in e-mail traffic," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 40, pp. 14333–14337, 2004.

[126] T. Fujisaka, R. Lee, and K. Sumiya, "Detection of unusually crowded places through micro-blogging sites," in *Advanced Information Networking and Applications Workshops (WAINA), 2010 IEEE 24th International Conference on*, pp. 467–472, IEEE, 2010.

[127] Y. Dong, F. Pinelli, Y. Gkoufas, Z. Nabi, F. Calabrese, and N. V. Chawla, "Inferring unusual crowd events from mobile phone call detail records," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 474–492, Springer, 2015.

[128] T. Aledavood, S. Lehmann, and J. Saramäki, "Digital daily cycles of individuals," *Frontiers in Physics*, vol. 3, 2015.

[129] J. Saramäki, E. A. Leicht, E. López, S. G. Roberts, F. Reed-Tsochas, and R. I. Dunbar, "Persistence of social signatures in human communication," *Proceedings of the National Academy of Sciences*, vol. 111, no. 3, pp. 942–947, 2014.

[130] G. Miritello, R. Lara, M. Cebrian, and E. Moro, "Limited communication capacity unveils strategies for human interaction," *Scientific reports*, vol. 3, 2013.

[131] T. Aledavood, E. López, S. G. Roberts, F. Reed-Tsochas, E. Moro, R. I. Dunbar, and J. Saramäki, "Channel-specific daily patterns in mobile phone communication," in *Proceedings of ECCS 2014*, pp. 209–218, Springer, 2016.

[132] T. Aledavood, E. López, S. G. Roberts, F. Reed-Tsochas, E. Moro, R. I. Dunbar, and J. Saramäki, "Daily rhythms in mobile telephone communication," *PloS one*, vol. 10, no. 9, p. e0138098, 2015.

[133] D. Minors and J. Waterhouse, *Circadian Rhythms and the Human*. Elsevier Science, 2013.

[134] R. Foster and L. Kreitzman, *Circadian Rhythms: A Very Short Introduction*. Very Short Introductions Series, Oxford University Press, 2017.

[135] O. Salvenmoser and B. Meklau, *Biological Clocks: Effects on Behavior, Health, and Outlook*. Public health in the 21st century series, Nova Science Publishers, 2010.

[136] J. Aschoff, "Exogenous and endogenous components in circadian rhythms," in *Cold Spring Harbor symposia on quantitative biology*, vol. 25, pp. 11–28, Cold Spring Harbor Laboratory Press, 1960.

[137] S. Daan and J. Aschoff, "The entrainment of circadian systems," *Handbook of behavioral neurobiology: Circadian clocks*, vol. 12, pp. 7–42, 2001.

[138] J. F. Duffy and K. P. Wright Jr, "Entrainment of the human circadian system by light," *Journal of biological rhythms*, vol. 20, no. 4, pp. 326–338, 2005.

[139] M. Melbin, "Night as frontier," *American Sociological Review*, pp. 3–22, 1978.

[140] S. Garbarino and L. Nobili, "Lifestyle and habits," in *Sleepiness and Human Impact Assessment*, pp. 95–103, Springer, 2014.

[141] O. Froy, "Circadian rhythms, aging, and life span in mammals," *Physiology*, vol. 26, no. 4, pp. 225–235, 2011.

[142] D.-J. Dijk and M. von Schantz, "Timing and consolidation of human sleep, wakefulness, and performance by a symphony of oscillators," *Journal of biological rhythms*, vol. 20, no. 4, pp. 279–290, 2005.

[143] A. Adan and V. Natale, "Gender differences in morningness–eveningness preference," *Chronobiology international*, vol. 19, no. 4, pp. 709–720, 2002.

[144] A. Adan, S. N. Archer, M. P. Hidalgo, L. Di Milia, V. Natale, and C. Randler, "Circadian typology: a comprehensive review," *Chronobiology international*, vol. 29, no. 9, pp. 1153–1175, 2012.

[145] I. Tsaousis, "Circadian preferences and personality traits: A meta-analysis," *European Journal of Personality*, vol. 24, no. 4, pp. 356–373, 2010.

[146] L. Kreitzman and R. Foster, *The Rhythms Of Life: The Biological Clocks That Control the Daily Lives of Every Living Thing*. Profile Books, 2011.

[147] T. Roenneberg, T. Kuehnle, M. Juda, T. Kantermann, K. Allebrandt, M. Gordijn, and M. Merrow, "Epidemiology of the human circadian clock," *Sleep medicine reviews*, vol. 11, no. 6, pp. 429–438, 2007.

[148] J. A. Horne and O. Östberg, "A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms.," *International journal of chronobiology*, vol. 4, no. 2, pp. 97–110, 1975.

[149] O. Östberg, "Interindividual differences in circadian fatigue patterns of shift workers," *Occupational and Environmental Medicine*, vol. 30, no. 4, pp. 341–351, 1973.

[150] M. Kryger, T. Roth, and W. Dement, *Principles and Practice of Sleep Medicine E-Book*. Elsevier Health Sciences, 2015.

[151] K. S. Jankowski, "Polish version of the reduced morningness–eveningness questionnaire," *Biological rhythm research*, vol. 44, no. 3, pp. 427–433, 2013.

[152] A. Adan and H. Almirall, "Horne & östberg morningness-eveningness questionnaire: A reduced scale," *Personality and Individual differences*, vol. 12, no. 3, pp. 241–253, 1991.

[153] R. Urbán, T. Magyaródi, and A. Rigó, "Morningness-eveningness, chronotypes and health-impairing behaviors in adolescents," *Chronobiology international*, vol. 28, no. 3, pp. 238–247, 2011.

[154] V. Natale, M. J. Esposito, M. Martoni, and M. Fabbri, "Validity of the reduced version of the morningness–eveningness questionnaire," *Sleep and biological rhythms*, vol. 4, no. 1, pp. 72–74, 2006.

[155] C. Randler, "German version of the reduced morningness–eveningness questionnaire (rmeq)," *Biological rhythm research*, vol. 44, no. 5, pp. 730–736, 2013.

[156] T. Roenneberg, A. Wirz-Justice, and M. Merrow, "Life between clocks: daily temporal patterns of human chronotypes," *Journal of biological rhythms*, vol. 18, no. 1, pp. 80–90, 2003.

[157] R. Levandovski, E. Sasso, and M. P. Hidalgo, "Chronotype: a review of the advances, limits and applicability of the main instruments used in the literature to assess human phenotype," *Trends in psychiatry and psychotherapy*, vol. 35, no. 1, pp. 3–11, 2013.

[158] S. Meliopoulos, S. Sheikh, J. Schneider, and R. Wattenhofer, "Analysis of sleeping patterns using smartphone sensors," *Semester Thesis*, 2011.

[159] Z. Chen, M. Lin, F. Chen, N. D. Lane, G. Cardone, R. Wang, T. Li, Y. Chen, T. Choudhury, and A. T. Campbell, "Unobtrusive sleep monitoring using smartphones," in *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2013 7th International Conference on*, pp. 145–152, IEEE, 2013.

[160] J.-K. Min, A. Doryab, J. Wiese, S. Amini, J. Zimmerman, and J. I. Hong, "Toss'n'turn: smartphone as sleep and sleep quality detector," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 477–486, ACM, 2014.

[161] A. Cuttone, P. Bækgaard, V. Sekara, H. Jonsson, J. E. Larsen, and S. Lehmann, "Sensiblesleep: a bayesian model for learning sleep patterns from smartphone events," *PloS one*, vol. 12, no. 1, p. e0169901, 2017.

[162] E. L. Murnane, S. Abdullah, M. Matthews, M. Kay, J. A. Kientz, T. Choudhury, G. Gay, and D. Cosley, "Mobile manifestations of alertness: connecting biological rhythms with patterns of smartphone app use.," in *MobileHCI*, pp. 465–477, 2016.

[163] E. K. Choe, S. Abdullah, M. Rabbi, E. Thomaz, D. A. Epstein, F. Cordeiro, M. Kay, G. D. Abowd, T. Choudhury, J. Fogarty, *et al.*, "Semi-automated tracking: A balanced approach for self-monitoring applications," *IEEE Pervasive Computing*, vol. 16, no. 1, pp. 74–84, 2017.

[164] S. Abdullah, M. Matthews, E. L. Murnane, G. Gay, and T. Choudhury, "Towards circadian computing: early to bed and early to rise makes some of us unhealthy and sleep deprived," in *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*, pp. 673–684, ACM, 2014.

[165] J. H. Lee, I. S. Kim, S. J. Kim, W. Wang, and J. F. Duffy, "Change in individual chronotype over a lifetime: a retrospective study," *Sleep Medicine Research (SMR)*, vol. 2, no. 2, pp. 48–53, 2011.

[166] M. Wittmann, J. Dinich, M. Merrow, and T. Roenneberg, "Social jetlag: misalignment of biological and social time," *Chronobiology international*, vol. 23, no. 1-2, pp. 497–509, 2006.

[167] T. Aledavood, S. Lehmann, and J. Saramäki, "Digital daily cycles of individuals." Submitted.

[168] Y. Selvi, A. Aydin, A. Atli, M. Boysan, F. Selvi, and L. Besiroglu, "Chronotype differences in suicidal behavior and impulsivity among suicide attempters," *Chronobiology international*, vol. 28, no. 2, pp. 170–175, 2011.

[169] C. Randler, "Association between morningness–eveningness and mental and physical health in adolescents," *Psychology, Health & Medicine*, vol. 16, no. 1, pp. 29–38, 2011.

[170] M. Wittmann, M. Paulus, and T. Roenneberg, "Decreased psychological well-being in late 'chronotypes' is mediated by smoking and alcohol consumption," *Substance use & misuse*, vol. 45, no. 1-2, pp. 15–30, 2010.

[171] H. Kontrymowicz-Ogińska, "Chronotype–behavioural aspects, personality correlates, health consequences," 2012.

[172] M. B. Horzum, İ. Önder, and Ş. Beşoluk, "Chronotype and academic achievement among online learning students," *Learning and Individual Differences*, vol. 30, pp. 106–111, 2014.

[173] A. L. D. Medeiros, D. B. Mendes, P. F. Lima, and J. F. Araujo, "The relationships between sleep-wake cycle and academic performance in medical students," *Biological Rhythm Research*, vol. 32, no. 2, pp. 263–270, 2001.

[174] J. Bonke, "Do morning-type people earn more than evening-type people? how chronotypes influence income," *Annals of Economics and Statistics/ANNALES D'ÉCONOMIE ET DE STATISTIQUE*, pp. 55–72, 2012.

[175] J. F. Duffy and C. A. Czeisler, "Age-related change in the relationship between circadian period, circadian phase, and diurnal preference in humans," *Neuroscience letters*, vol. 318, no. 3, pp. 117–120, 2002.

[176] I. Tankova, A. Adan, and G. Buela-Casal, "Circadian typology and individual differences. a review," *Personality and individual differences*, vol. 16, no. 5, pp. 671–684, 1994.

[177] R. Ling, ""she calls, [but] it's for both of us you know"," *The use of traditional fixed and mobile telephony for social networking among Norwegian parents. Kjeller: Telenor.(R & D. Report 33/98-2000)*, 1998.

[178] V. Frissen *et al.*, "Gender is calling: Some reflections on past, present and future uses of the telephone," *The gender-technology relation: Contemporary theory and research*, pp. 79–94, 1995.

[179] J.-P. Onnela, B. N. Waber, A. Pentland, S. Schnorf, and D. Lazer, "Using sociometers to quantify social interaction patterns," *Scientific reports*, vol. 4, p. 5604, 2014.

[180] V. Frias-Martinez, E. Frias-Martinez, and N. Oliver, "A gender-centric analysis of calling behavior in a developing economy using call detail records.," in *AAAI spring symposium: artificial intelligence for development*, 2010.

[181] V. Palchykov, K. Kaski, J. Kertész, A.-L. Barabási, and R. I. Dunbar, "Sex differences in intimate relationships," *Scientific reports*, vol. 2, p. 370, 2012.

[182] N. Eagle, M. Macy, and R. Claxton, "Network diversity and economic development," *Science*, vol. 328, no. 5981, pp. 1029–1031, 2010.

[183] "Sociopatterns." http://www.sociopatterns.org.

[184] M. Starnini, B. Lepri, A. Baronchelli, A. Barrat, C. Cattuto, and R. Pastor-Satorras, "Robust modeling of human contact networks across different scales and proximity-sensing techniques," *arXiv preprint arXiv:1707.06632*, 2017.

[185] M. J. Keeling and K. T. Eames, "Networks and epidemic models," *Journal of the Royal Society Interface*, vol. 2, no. 4, pp. 295–307, 2005.

[186] A. Barrat, M. Barthélemy, and A. Vespignani, *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008.

[187] J. Read, W. Edmunds, S. Riley, J. Lessler, and D. Cummings, "Close encounters of the infectious kind: methods to measure social mixing behaviour," *Epidemiology & Infection*, vol. 140, no. 12, pp. 2117–2130, 2012.

[188] 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*, vol. 107, no. 51, pp. 22020–22025, 2010.

[189] T. Smieszek, S. Castell, A. Barrat, C. Cattuto, P. J. White, and G. Krause, "Contact diaries versus wearable proximity sensors in measuring contact patterns at a conference: method comparison and participants' attitudes," *BMC infectious diseases*, vol. 16, no. 1, p. 341, 2016.

[190] L. Isella, M. Romano, A. Barrat, C. Cattuto, V. Colizza, W. Van den Broeck, F. Gesualdo, E. Pandolfi, L. Ravà, C. Rizzo, *et al.*, "Close encounters in a pediatric ward: measuring face-to-face proximity and mixing patterns with wearable sensors," *PloS one*, vol. 6, no. 2, p. e17144, 2011.

[191] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. Van den Broeck, "What's in a crowd? analysis of face-to-face behavioral networks," *Journal of theoretical biology*, vol. 271, no. 1, pp. 166–180, 2011.

[192] J. Stehlé, F. Charbonnier, T. Picard, C. Cattuto, and A. Barrat, "Gender homophily from spatial behavior in a primary school: a sociometric study," *Social Networks*, vol. 35, no. 4, pp. 604–613, 2013.

[193] J. Fournet and A. Barrat, "Contact patterns among high school students," *PloS one*, vol. 9, no. 9, p. e107878, 2014.

[194] J. Torous, M. V. Kiang, J. Lorme, and J.-P. Onnela, "New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research," *JMIR mental health*, vol. 3, no. 2, 2016.

[195] J. Torous, M. Keshavan, J.-p. Onnela, P. Staples, and I. Barnett, "M48. digital phenotyping in schizophrenia using smartphones," *Schizophrenia Bulletin*, vol. 43, no. suppl_1, p. S228, 2017.

[196] J. Torous, P. Staples, L. Sandoval, I. Barnett, J. P. Onnela, and M. Keshavan, "323. utilizing smartphones to collect longitudinal digital phenotypes in patients with schizophrenia," *Biological Psychiatry*, vol. 81, no. 10, pp. S132–S133, 2017.

[197] "Beiwe platform." https://www.hsph.harvard.edu/onnela-lab/beiwe-research-platform/. Accessed: 2017-08-11.

[198] D. Ferreira, V. Kostakos, and A. K. Dey, "Aware: mobile context instrumentation framework," *Frontiers in ICT*, vol. 2, p. 6, 2015. doi:10.3389/fict.2015.00006.

[199] D. Ferreira, *AWARE: A Mobile Context Instrumentation Middleware To Collaboratively Understand Human Behavior*. PhD thesis, University of Oulu, 2013.

[200] A. K. Dey, K. Wac, D. Ferreira, K. Tassini, J.-H. Hong, and J. Ramos, "Getting closer: an empirical investigation of the proximity of user to their smart phones.," in *Ubicomp'11*, pp. 163–172, 2011.

[201] J. Goncalves, D. Ferreira, S. Hosio, Y. Liu, J. Rogstadius, H. Kukka, and V. Kostakos, "Crowdsourcing on the spot: altruistic use of public displays, feasibility, performance, and behaviours," in *Ubicomp'13*, pp. 753–762, ACM, 2013.

[202] "Aware framework." http://www.awareframework.com.

[203] B. Mittelstadt and L. Floridi, *The Ethics of Biomedical Big Data*. Law, Governance and Technology Series, Springer International Publishing, 2016.

[204] S. D. Warren and L. D. Brandeis, "The right to privacy," *Harvard law review*, pp. 193–220, 1890.

[205] J. A. Barnes, "Who should know what?: social science, privacy, and ethics,", 1979.

[206] O. Tene and J. Polonetsky, "Privacy in the age of big data: a time for big decisions," *Stan. L. Rev. Online*, vol. 64, p. 63, 2011.

[207] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *ACM Sigmod Record*, vol. 29, pp. 439–450, ACM, 2000.

[208] A. L. Young and A. Quan-Haase, "Privacy protection strategies on facebook: The internet privacy paradox revisited," *Information, Communication & Society*, vol. 16, no. 4, pp. 479–500, 2013.

[209] G. A. of the World Medical Association *et al.*, "World medical association declaration of helsinki: ethical principles for medical research involving human subjects.," *The Journal of the American College of Dentists*, vol. 81, no. 3, p. 14, 2014.

[210] J. Cohen and T. Ezer, "Human rights in patient care: A theoretical and practical framework," *health and human rights*, vol. 15, no. 2, pp. 7–19, 2013.

[211] J. Wang, Y. Luo, Y. Zhao, and J. Le, "A survey on privacy preserving data mining," in *Database Technology and Applications, 2009 First International Workshop on*, pp. 111–114, IEEE, 2009.

[212] European Parliament, Council of the European Union, "Regulation (EU) 2016/679, Official Journal of 4 May 2016, L 119, p. 1–88," 2016.

[213] J. M. M. Rumbold and B. Pierscionek, "The effect of the general data protection regulation on medical research," *Journal of medical Internet research*, vol. 19, no. 2, 2017.

[214] S. Ji, P. Mittal, and R. Beyah, "Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: A survey," *IEEE Communications Surveys & Tutorials*, 2016.

[215] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, SP '08, (Washington, DC, USA), pp. 111–125, IEEE Computer Society, 2008.

[216] A. Singh, D. Bansal, and S. Sofat, "Privacy preserving techniques in social networks data publishing-a review," *International Journal of Computer Applications*, vol. 87, no. 15, 2014.

[217] "Koota server." https://github.com/CxAalto/koota-server.

[218] A. Monreale, S. Rinzivillo, F. Pratesi, F. Giannotti, and D. Pedreschi, "Privacy-by-design in big data analytics and social mining," *EPJ Data Science*, vol. 3, no. 1, p. 10, 2014.

[219] P. Y. Collins, V. Patel, S. S. Joestl, D. March, T. R. Insel, A. S. Daar, I. A. Bordin, E. J. Costello, M. Durkin, C. Fairburn, *et al.*, "Grand challenges in global mental health," *Nature*, vol. 475, no. 7354, pp. 27–30, 2011. PMID:21734685.

[220] P. Boksa, "A way forward for research on biomarkers for psychiatric disorders," *J Psychiatry Neurosci*, vol. 38, no. 2, pp. 75–7, 2013. doi: 10.1503/jpn.130018.

[221] S. Shiffman, A. A. Stone, and M. R. Hufford, "Ecological momentary assessment," *Annu. Rev. Clin. Psychol.*, vol. 4, pp. 1–32, 2008.

[222] A. A. Stone and S. Shiffman, "Ecological momentary assessment (ema) in behavorial medicine.," *Annals of Behavioral Medicine*, 1994.

[223] S. H. Jain, B. W. Powers, J. B. Hawkins, and J. S. Brownstein, "The digital phenotype," *Nature biotechnology*, vol. 33, no. 5, pp. 462–463, 2015.

[224] D. Ben-Zeev, C. J. Brenner, M. Begale, J. Duffecy, D. C. Mohr, and K. T. Mueser, "Feasibility, acceptability, and preliminary efficacy of a smartphone intervention for schizophrenia," *Schizophrenia bulletin*, p. sbu033, 2014. PMID: 24609454.

[225] M. Faurholt-Jepsen, M. Vinberg, E. M. Christensen, M. Frost, J. Bardram, and L. V. Kessing, "Daily electronic self-monitoring of subjective and objective symptoms in bipolar disorder—the monarca trial protocol (monitoring, treatment and prediction of bipolar disorder episodes): a randomised controlled single-blind trial," *BMJ open*, vol. 3, no. 7, p. e003353, 2013. PMID: 23883891.

[226] J. Torous, P. Staples, M. Shanahan, C. Lin, P. Peck, M. Keshavan, and J.-P. Onnela, "Utilizing a personal smartphone custom app to assess the patient health questionnaire-9 (phq-9) depressive symptoms in patients with major depressive disorder," *JMIR mental health*, vol. 2, no. 1, 2015. PMID:26543914.

Aalto-DD 225/2017

ISBN 978-952-60-7723-9 (printed)
ISBN 978-952-60-7724-6 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Computer Science
www.aalto.fi

BUSINESS +
ECONOMY

ART +
DESIGN +
ARCHITECTURE

SCIENCE +
TECHNOLOGY

CROSSOVER

DOCTORAL
DISSERTATIONS