Multichannel Social Signatures and Persistent Features of Egocentric Networks

Sara Heydari

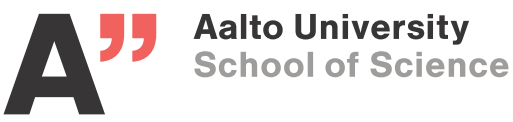
School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 17.04.2018

Thesis supervisor:

Prof. Jari Saramäki



AALTO UNIVERSITY
SCHOOL OF SCIENCE

ABSTRACT OF THE
MASTER'S THESIS

Author: Sara Heydari

Title: Multichannel Social Signatures and Persistent Features of Egocentric Networks

Date: 17.04.2018 Language: English Number of pages: 6+45

Department of Computer Science

Professorship: Computational Science

Supervisor and advisor: Prof. Jari Saramäki

Mobile phones are perfect sensors for capturing the behavior of people. They are widespread personal devices that we carry around all day. Modern smartphones, equipped with an arsenal of various sensors, monitor their environments and also their owners. However, even the simplest mobile phone device, when used with a SIM card, can collect rich behavioral data. Call Detail Records (CDRs), collected by telecommunication companies for billing purposes, contain detailed information on communication behavior of the users which can not be collected by traditional data collection methods such as questionnaires. Scientists have used CDRs to study the structure and dynamics of societal-level communication networks as well as the properties of egocentric networks. The structure of weighted egocentric networks can be quantified with the so-called *social signatures*. It is known that call-based social signatures are distinct and persistent at the individual level. However, calling is just one of the several channels that people use to communicate. To get a more realistic picture of people's social behavior we should include more communication channels. However, because of their intrinsic differences, it is challenging to combine the usage frequencies on multiple channels into single combined weights. In this Thesis, we propose a method for determining link weights which enables us to compare the egocentric networks across different channels and also to construct multichannel egocentric networks and multichannel social signatures. Using two different datasets on calling and texting behavior of people, we observed that similarly to call signatures, text-message signatures and multichannel signatures (combining information on calls and texts) are also persistent in time. Moreover, we observed that even though people call and text different sets of people, their call and text signatures are similar in shape. In other words, the shapes of our social signatures–which are distinct from signatures of others–seem to be independent of the communication channel or the people whom we contact. Further research is needed to explain the mechanism behind these shapes and to investigate the roots of persistence and stability of social signatures.

Keywords: Social Networks, Social Signatures, Egocentric Networks, Inferring Social Ties

Preface

I began my Master's studies in the autumn of year 2015. Since the very beginning, I got the chance to learn and do research in the complex systems research group of Aalto University. This Thesis contains the key results I found during these years. Now I am at the end of the path that I started almost three years ago when I left my hometown, my beloved family and friends to pursue my passion.

First and foremost, I want to thank my supervisor, Jari Saramäki for believing in me even at the moments that I did not believe in myself. I feel so lucky for having such a supportive and ingenious supervisor.

I am also grateful to all my colleges in our research group for their company and providing their help whenever I needed and for all the coffee breaks, fruit hours, movie nights and interesting conversations that we had. You have taught me a lot and made the work more enjoyable for me: Mikko, Tuomas, Arash, Talayeh, Ana, Onerva, Ilkka, Rainer, Richard, Darko, Gerardo, Christofer, Sallamari, Kimmo, Pietro, Daniel, Kunal, and Asim.

I wish also to thank all those people who gave me the joy of learning. Among them are Prof. Jafari who made me interested in complex systems and supported me to carry on in this field and Prof. Sepanji who was always there for us students.

I also want to thank all my friends around the world who brought the light to me in the dark nights.

Finally, I wish to thank my parents for giving me the freedom and courage to find my own way. Your love gave me the strength throughout my journey.

Tehran, 14.04.2018

Sara Heydari

Contents

Abstract ii

Preface iii

Contents iv

Abbreviations vi

1 Introduction 1

2 Background 4
2.1 A Story of Humans: From the Stone Age to the Information Age . . . 4
2.2 Computational Social Science . . . 6
2.3 Mobile Phone Datasets . . . 6
2.3.1 Mobile Phone Call Detail Records . . . 7
2.3.2 Mobile Phone Data Collection Studies . . . 7
2.4 Network Science . . . 8
2.5 Social networks . . . 9
2.5.1 Six Degrees of Separation . . . 10
2.5.2 The Watts-Strogatz Small-World Model . . . 10
2.5.3 The Barabási–Albert Model . . . 11
2.5.4 The Strength of Weak ties . . . 12
2.5.5 Egocentric Networks . . . 13
2.5.6 Burstiness in Human Communication . . . 13
2.5.7 Challenges in Detection of Social Ties and Measuring Their Strength . . . 15

3 Datasets 17
3.1 The large mobile phone dataset (DS1) . . . 17
3.1.1 Basic Preprocessing . . . 19
3.1.2 Filtering Based on Activity Level . . . 19
3.1.3 Data Statistics Among Active Users . . . 20
3.2 The UK Students Dataset (DS2) . . . 21

4 Methods and Results 23
4.1 Defining Egocentric Networks . . . 23
4.2 Social Signatures . . . 24
4.3 Comparing Social Signatures . . . 26
4.3.1 L^2 Distance . . . 26
4.3.2 Jensen-Shannon Distance . . . 27
4.3.3 Jaccard Index . . . 27
4.4 Results . . . 28
4.4.1 Call, Text and Mixed Social Signatures Are Persistent in Time 28

4.4.2 Call and Text Signatures Have Similar Shapes, Both at the Population and at the Ego Levels 31
4.4.3 Call and Text Egocentric Networks Differ in Composition . . . 31
4.4.4 Composition of Mixed Social Signatures: Channel Choice Does Not Depend on Alter Rank 34

5 Summary and Discussion **37**

References **39**

Abbreviations

Operators

$\sum_i$	sum over index i
$H(s_1)$	Shannon entropy of S_1

Abbreviations

CDR	call detail record
ENIAC	Electronic Numerical Integrator and Computer
BRLESC	Ballistic Research Laboratories Electronic Scientific Computer
WWW, the Web	the World Wide Web
RFID	radio-frequency identification
GPS	global Positioning System
SIM card	subscriber identity module card
MIT	Massachusetts Institute of Technology
B-A model	Barabási–Albert model
SMS	short message service
DS1	Dataset1 (the large mobile phone dataset)
DS2	Dataset2 (the UK students dataset)
ID	identifier
L^2 distance	Euclidean distance
JSD	Jensen-Shannon divergence
KLD	Kullback-Leibler divergence
$JS_{distance}$	Jensen-Shannon distance

1 Introduction

Social relationships are crucial for the well-being of humans and also a necessity for the human society. Research has shown that emotionally intensive relationships are essential not only to the health of humans but also to that of other primates [1, 2, 3]. On the other hand, less intensive relationships (weaker ties) bring us diversity, leading out of our everyday social circles and providing us novel information and opportunities [4].

However, maintaining social relationships is costly and we have limited resources. To retain emotional closeness in social relationships, frequent communication is needed, and this is time-consuming [5, 6]. Obviously, we – mortal humans – have limited time. Moreover, there are studies showing that our brain capacity is another limitation confining our social lives [7, 8, 9]. In particular, maintaining close relationships is exceedingly time-consuming and also cognitively demanding [10]. As a result, our personal social networks usually comprise a few close ties and numerous weak ties. This behavioral characteristic of humans impacts the dynamics of phenomena on society-wide communication networks [11] along with the structure of egocentric networks [12].

In 2013, Refs. [13] and [12] reported that people distribute their mobile phone calls unevenly among their alters: a few close alters receive a disproportionally large fraction of calls while the rest is divided among a large number of alters. Ref. [12] suggested that this disparity in the distribution of communication can be quantified with so-called *"social signatures"*. A social signature measures the fraction of communication devoted to each alter as a function of their rank, when the alters are ranked based on the communication fraction. The important observation reported in Ref.[12] was that each individual has a distinct social signature which persists in time despite turnover in her personal network. Later, another study also reported persistence of call-based social signatures using a different dataset [14]. Moreover, in 2015, Ref. [15] reported similar results on social signatures constructed from email networks.

In this Thesis, we study the structure of personal networks–so-called egocentric networks–and properties of social signatures based on information on communication through two different channels: mobile telephone calls and text messages. To do so, we use two different call detail record (CDR) datasets dating back to 2007 and 2008. In those years, mobile phone subscribers used calling and texting services provided by mobile phone operators more exclusively than today, where much of communication is done using different smartphone applications. One of the datasets is small in size (24 subjects) but detailed and carefully collected and the other is a large dataset containing the communication data of more than half a million users over 7 months. The datasets are explained in detail in Sec. 3.

As already mentioned, the communication datasets used in here contain information on mobile-phone calls and text messages. The two communication channels of calls and text messages are intrinsically different. In general, humans use a wide and diverse range of channels to maintain their social relationships. These channels have different features and functionality. Thus the person who initiates a communication

event does not pick a channel randomly, but chooses the appropriate channel based on numerous factors including purpose of communication, type of relationship, general channel preferences of the initiator, and the time of communication (morning/evening, weekday/weekend).

As a result, in order to get a more realistic picture of egocentric networks and social behavior of individuals, we need to include information on multiple communication channels. However, this is challenging, first of all simply because we usually do not have access to such comprehensive datasets. Moreover, because of the intrinsic differences between the channels, it is hard to compare and combine information on communication through different channels. In this Thesis, we first propose a method which by using the timestamps of communication events enables us to define comparable link weights for call-based and text-based egocentric networks or even construct a combined egocentric network (see Sec. 4.1).

Then, equipped with comparable egocentric networks, in Sec. 4.2, we construct call, text, and mixed social signatures of individuals. It has been reported earlier in the literature that individuals have distinct and persistent phone call and email social signatures [12, 14, 15]. Here, using data on communication behavior of a large and diverse sample of more than half a million people, we also observe that each individual has a distinct phone call social signature compared to others, which is persistent in time. Moreover, we show that text message and mixed social signatures also have the same property.

Further, in Sec. 4.4.2, we compare call social signatures with text message social signatures. We observe that call and text signatures of an individual during the same period of time are similar in shape when compared to the social signatures of others. This result is surprising, because the call network and text network of a person are usually very different in membership. We do not text the same set of alters that we call and moreover, those alters that we contact by both phone calls and text messages usually acquire different ranks in our call and text social signatures.

To find an explanation for the similarity of call and text social signatures, in Sec. 4.4.4, we take a closer look at the composition of mixed social signatures. For instance, if share of call communication would be the same for all alters, the call and text signatures would consequently be identical. However, we observe that the choice of channel is neither regular nor predictable from the alters' rank.

To conclude, several studies on datasets from different countries and different communication channels agree that individuals have distinct social signatures which are persistent in time [12, 14, 15]. Here, we observed that individuals also have similar-looking signatures across different channels of communication. Why do individuals then have social signatures which persist in time and have similar shapes across different channels? Further studies are needed to answer this question. One can hypothesize that each ego has a latent underlying egocentric network which determines the strength of all the ego's relationships. Then, each single-channel social signature is made from a sample of the ego's social actions and reflects an incomplete picture of the latent overall social signature. Then the question would be how the interactions via different channels sample the underlying network so that single-channel social signatures–these incomplete reflections of the latent signature–end up

being similar in shape.

This Master's Thesis is structured as follows: in Chapter 2, I provide a background on computational social science, the use of mobile phone datasets in this field, network science as a tool to study social networks, egocentric networks, and measuring tie strength from electronic records. In Chapter 3, I explain the two datasets used in this Thesis in detail and report their statistics. Then, in Chapter 4, I explain the methods applied in this Thesis and report our observations and results. Finally, in the last chapter, I provide a summary of the Thesis and discuss ideas for future studies.

2 Background

2.1 A Story of Humans: From the Stone Age to the Information Age

Around 12,000 years ago, the hunter-gatherer humans went through a transition in their lifestyle which is referred to as the agricultural revolution. The sedentary lifestyle which itself became possible as a result of the domestication of plants and animals gave humankind the opportunity to observe, test and learn how to produce more food. The availability of surplus food made the increase of population possible and the settlements expanded [16]. The development of densely populated settlements gave birth to the early civilizations.

The industrial revolution was the next turning point in the history of the modern human. The transition from manual production methods to machine production happened during the 18th and early 19th centuries. As a consequence of the increase in production, the human population and wealth grew vastly and once again the human-kind experienced a lifestyle change: the industrial revolution influenced almost every aspect of people's daily lives [17].

The technological revolution, or the second industrial revolution, took place from 1870 to the beginning of the First World War. Railroads and telegraph networks became widespread which eased the movement of people as well as the exchange of information and ideas. Moreover, during that era, people managed to harness the power of electricity. The world's first modern power station was built in 1891. Also, the telephone and the radio were invented which made people more connected [18].

In the 20th century, the world wars expedited technological advances. Among the most important advances were developments in electronics, computers and computational methods. In 1947, ENIAC, one of the earliest programmable computers, was used to study explosion of a thermonuclear weapon [19, 20]. The computer program simulating the process was the first program using pseudo-random numbers–which are still used widely in programming nowadays. ENIAC, a pioneer of computational power in its own time, weighted around 30 tons and used punch cards to receive input and return output [21].

The invention of transistors in 1947, facilitated the development of digital computers [22]. This was beginning of another transition in human history: *the digital revolution*. Since then, computers have shrunk extremely in size and their computational power has increased exponentially [23]. As a result of miniaturization, the desktop computers which were popular in the 80's had more computational power than the 30-ton ENIAC.

The invention of World Wide Web (the Web) in 1989 along with ubiquitousity of small and powerful computers was a focal point in the transition to the *information age*: computers were not anymore merely computational machines, but they were increasingly used also for communication and exchange of information.

The digital revolution, similar to the agricultural and industrial revolutions, has changed the lifestyle of humankind. We are citizens of the information age and our routines reflect this fact. The day of a person in our time can typically start like this:

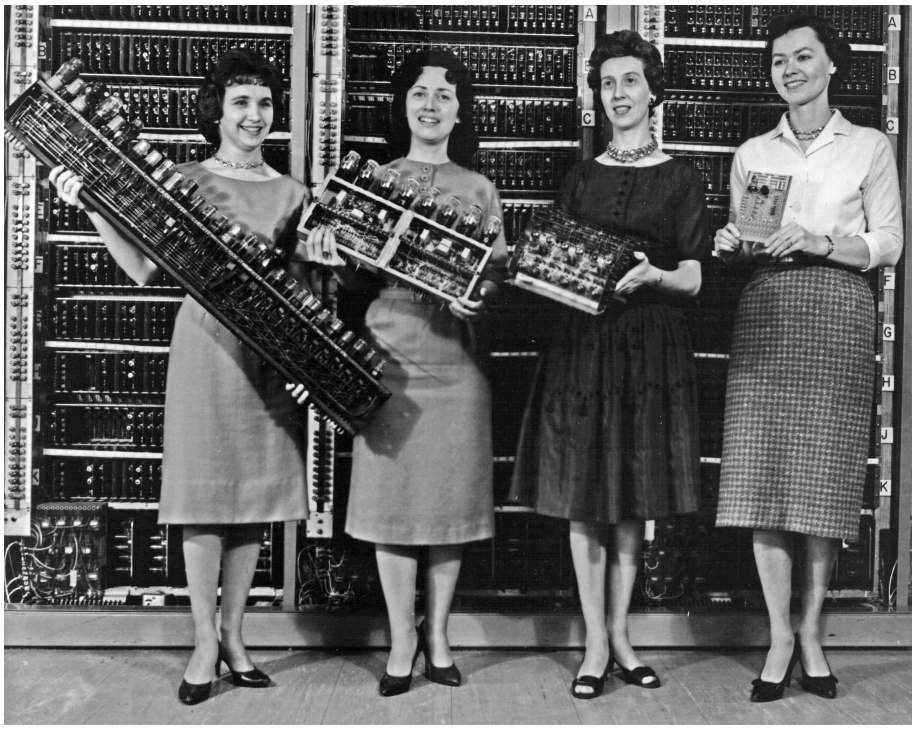


Figure 1: "U.S. Army Photo", number 163-12-62. The photo, taken in 1962, shows the miniaturization trend perfectly. The women are holding boards of the first four U.S. army computers. The leftmost board belongs to ENIAC which started to work in 1946, and the rightmost board belongs to BRLESC-I which began operation in 1962.

She wakes up with the alarm set on her smartphone – the powerful tiny computer that she carries the whole day in her pocket. Then she answers a few messages on her phone while still lying in bed. Then, while eating breakfast, she checks her online calendar on her phone for the day's agenda, pays the electricity bill, and also checks when the next bus leaves. Then she gets on the bus paying the ticket by a RFID travel card. While sitting on the bus, she checks her Twitter timeline, reading about the presidential election in France and also getting to know that her high-school friend's daughter has won a skating competition. While doing all these ordinary, everyday activities, we produce an enormous amount of data. Almost whatever we do, we leave behind "digital breadcrumbs" [24].

However, despite all these modern social technologies, our brains are still more or less those of hunter-gatherers. It has, in fact, been argued that much of our social behavior still reflects the environment where the early humans evolved; e.g. our personal network size which is limited by our brain capacity is thought to reflect the size of a proto-human tribe [25].

2.2 Computational Social Science

The existence of the massive amount of quantitative data on the behavior of humans and the availability of gigantic computational power of modern computers for analyzing these data have attracted natural scientists–especially physicists and computer scientists–to apply their methods to study the social behavior of people. This has given rise to the new interdisciplinary field of *computational social science* [26, 27, 24].

As opposed to traditional social sciences, which are fueled by time-consuming and expensive field observations and surveys [28], computational social science relies on passively collected data on the behavior of people. These datasets can be found in various forms such as email records, social media logs, location data collected by GPS, or call detail records collected by teleoperators for billing purposes. These datasets are typically anonymized to protect the privacy of subjects.

An advantage of these electronic records over traditional labor-intensive surveys is their scalability. These electronic records can contain data on the social interactions and behavior of millions of people [27]. Therefore, they enable us to study the structure and dynamics of social systems at the societal level [11]. Moreover, questionnaire-based datasets limit us to what people can recall. It is known that human recall of the past events is not accurate [29]. Passively recorded datasets, on the other hand, can provide a continuous and detailed picture of human behavior. For example, they enable scientists to study details such as exact timings of communication events which are impossible to collect by questionnaires. Such high resolution and extended datasets, for instance, make it possible to study temporal patterns and causalities of human behavior [27, 30, 31, 32].

Computational social science is not just an alternative to the traditional social sciences, it is the future of social sciences. A paradigm shift from traditional social sciences to computational social science is happening [33]. The emergence of quantitative high-resolution data on the behavior of millions of people brings us the opportunity to understand and predict human society in a way that was not possible two decades ago. However, this is not feasible unless we utilize proper tools and methods to harness, analyze and explore these datasets.

2.3 Mobile Phone Datasets

Mobile phones are perfect sensors for capturing the behavior of human individual for several reasons. First, they are widespread: in 2014, the number of mobile phone subscriptions was about 96% of the world's population [34]. Second, they are usually used as a personal device, so the data collected by them presents the behavioral patterns of one single individual. Moreover, we keep them in our pockets and carry them around, so they can be used to track our behavior closely and with a high resolution. There are two main types of mobile phone datasets: call detail records datasets (CDRs) collected by mobile operators for billing reasons, and datasets collected by mobile phone applications installed on smart-phones which are usually specifically designed to collect behavioral data of the mobile phone holders. In here,

we briefly explain these two dataset types and review a few studies conducted on them.

2.3.1 Mobile Phone Call Detail Records

Nowadays, smartphones are equipped with an arsenal of various sensors which are monitoring their environments and also the mobile phone holders. However, even a basic and old-fashioned phone, if used with a SIM card, can collect rich data on the behavior of the user. Telecommunication companies keep a record of their customers' service usage for billing purposes. These records are referred to as Call Detail Records (CDRs) and contain all the information needed for billing purposes: the event type (Phone call, text message, multimedia message), the timestamp of the event, the duration of the call, and the location of the cell tower that the customer's phone has been connected to during the communication event. Moreover, sometimes the CDR datasets provided by operators contain some additional basic information such as the age and gender of customers. In the recent years, CDR datasets have been explored in a wide range of studies. A 2015 paper, Ref. [34], gives us a comprehensive review of this field of study.

As mentioned before, CDR datasets can be massive in sample size, so they allow us to study the structure of social networks at the societal level. Refs. [11] and [35] have used a CDR dataset containing 20% of the population of a country to study the structure of the social network made based on it. Besides being large in scale, CDR datasets include timings of the social events with a high resolution (with the precision of seconds) and are usually extended over rather long periods of time. As a result, they have been used to study the dynamics of spreading processes [36, 37, 38, 39] along with temporal patterns at the level of individuals [27, 40, 41, 42].

Moreover, the geographical information in CDR datasets makes it possible to use them to study the mobility of people. Ref. [43] studied the mobility patterns of individuals and observed that the patterns are highly regular. This indicates that models such as the Lévy flight are not a good approximation for the movement of humans. There have been several studies discussing predictability of human location from their past CDR information and trying to give a realistic model of the mobility of people [44, 45]. Moreover, CDR datasets, because of their high temporal resolution and massive size, provide the possibility to study the collective mobility patterns of humans, for example calculating travel times between cities [46] or extracting the tempo-geographical mobility patterns at the scale of the whole population of a city or a country [47, 48, 49].

2.3.2 Mobile Phone Data Collection Studies

CDRs are rich data sources for studying behavioral patterns of people, specifically because of their enormous size. However, they lack depth compared to data which can be collected using the variety of sensors embedded in smartphones. Moreover, by technological improvement of mobile phones, an increasingly smaller fraction of the users' communication is done via call and messaging services of teleoperators.

These reasons have motived scientists to design studies to collect behavioral data from smartphone users.

The MIT "Reality Mining" experiment which was run in 2004 is probably the first example of these data collection studies [50]. In that study, the behavior of one hundred MIT students and employees was tracked using Nokia 6600 phones. The devices were enhanced with a mobile application designed by scientists at the University of Helsinki specifically to collect behavioral data of the users [51]. The information collected by the devices includes call logs, cell tower IDs, information on nearby devices detected by the Bluetooth sensor, application usage and phone status (such as in use, idle, and charging). Moreover, the dataset was extended by survey data on some basic information about subjects, their friendships, and their lifestyles.

The "SensibleDTU" project, started in Technical University of Denmark in 2012, is a more recent example of mobile phone data collection studies. In the SensibleDTU, similarly to the Reality Mining experiment, the data is collected by identical mobile phone devices distributed among the participants and also by questionnaires filled by them. The smartphones were embedded by an application designed to collect the users' behavioral data. Because of technological advances, the smartphones used in this experiment, compared to devices used in the Reality Mining, were capable of collecting more detailed and diverse data. The collected data include location and proximity data via GPS, Bluetooth, WiFi, call logs and information on the participants' activity on the online social media Facebook [52].

To summarize, I should mention that the mobile phone data collection studies are limited in the number of participants compared to CDR datasets, but contain more details. Moreover, because of the increasing popularity of communication applications among smartphone users, CDRs are becoming less and less informative about people's social lives. As a result, there is increasingly more need to conduct data collection studies to reveal behavioral and social patterns of humans.

2.4 Network Science

Network science, the science of modeling real-world systems as networks and studying their properties, is an interdisciplinary field both in its origin and in its application areas. A network, referred to as "graph" in mathematics, is a set of objects in which some of the object pairs are related to each other in some way. In different disciplines, there are different terms for referring to the "objects" and the "relations": vertices and edges in graph theory, actors and ties in social sciences, sites and bonds in chemistry and physics, and nodes and links in computer science. The existence of various terminologies for referring to the same concepts indicates that different disciplines have contributed to the emergence of network science and that it has been used as a toolkit in different fields [19].

Computational Social Science is among the fields in which network methods are frequently used. In the next section, we discuss early examples of using networks in sociology and then explain some of the key characteristics of social networks. Wherever we talk about a concept in network science, we briefly explain it.

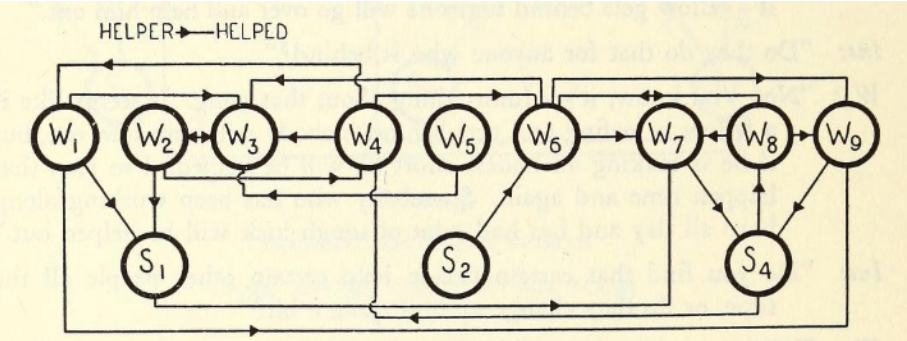


FIGURE 42
PARTICIPATION IN HELPING

Figure 2: A sociogram on helping relationships among workers drawn by Fritz Roethlisberger and published in his book titled *Management and the Worker* [56]. The book contains dozens of sociograms showing different types of relationships among workers and managers based on Roethlisberger's observations.

2.5 Social networks

Many concepts in the network science have their roots in graph theory, a field of mathematics initiated in 1713, when Leonhard Euler published a paper addressing the "Seven Bridges of Königsberg" problem [53]. However, it was around the end of the 19th century that graph-theoretical concepts started to be employed by scientists in other disciplines. Sociology was among the first disciplines to borrow the concepts of graph theory. In 1933, Jacob Moreno, a psychiatrist who is known as the founder of social network analysis, presented his studies on the dynamics of social interaction. A year later, he published a book titled *Who Shall Survive?* [54] which is considered to be the first published work in social network analysis [55].

Moreno, to visualize the relationships in groups of people, drew diagrams illustrating people as geometrical objects (e.g. circles) and the relationships between them as lines. This is how we still visualize networks. He called these diagrams *sociograms*. His simple and informative sociograms encouraged other sociologists to also use them. For example, Fritz Roethlisberger used sociograms in his 1939 book titled *Management and the Worker*, which was a study on relationships among managers and workers in a company [56] (see Fig. 2).

The early sociograms were drawn based on the data collected by field observations or surveys. Thus they were small in size. It was at the end of the 20th century that the availability of electronic records on the social behavior of people made it possible to study the properties of large-scale social networks. To understand how these large-scale real-world networks are originated and how they can evolve, network scientists started to develop graph-generating models which attempted to reproduce graphs similar to real-world networks [57]. These graph-generating models, like the Small-World model (1998) [58] and the Barabási-Albert model (1999) [59] were indeed inspired by the work of two mathematicians, Paul Erdös and Alfréd Rényi, on random graph generation models (1959-1961) [60].

2.5.1 Six Degrees of Separation

"Let us start with familiar observations: the 'small world' phenomenon, and the use of friends in high places to gain favors. It is almost too banal to cite one's favorite unlikely discovery of a shared acquaintance, which usually ends with the exclamation 'My, it's a small world!'." This is how the paper titled "Contacts and Influence" [61], authored around 1960 by the social scientist Ithiel de Sola Pool and the mathematician Manfred Kochen starts. Inspired by this paper, Stanley Milgram in 1967 conducted an experiment, Ref. [62], to estimate the average *distance* between pairs of nodes in the social network of people living in the United States [63]. The distance between two nodes in a network is defined as the number of links in the shortest path between the two nodes.

Milgram, in his experiment, picked hundreds of randomly selected people living on the West Coast and gave each of them a letter addressed to random recipients living on the East Coast. Each person was instructed to forward the letter directly if they knew the recipient personally, and otherwise send it to a friend or relative who was more likely to know the recipient. Based on this experiment, Milgram reported the average shortest path to be around five and a half. In 1990, the American playwright John Guare, inspired by the result of Milgram experiment, wrote a popular play titled "six degrees of separation" [64]. From then on, this phrase has been used many times to refer to Milgram's work or the small-world characteristic of social networks [63].

2.5.2 The Watts-Strogatz Small-World Model

In Erdös-Rényi random graphs, the probability of any two random nodes being connected is equal to p. The graphs generated by this model have one of the small world properties: even in the large graphs, the average shortest path is small. To be exact, the average shortest path is in the order of $\log N$, where N is the total number of the nodes in the graph.

Even though Erdös-Rényi graphs have small average shortest path length, the model is inappropriate for modeling real-world networks. This is primarily because the assumption that all pairs of nodes are equally probable to be connected is not realistic. Moreover, Erdös-Rényi graphs have a low *clustering coefficient* unlike many real-world networks, in particular social networks. The clustering coefficient of a node measures how probably the *neighbors* of a node are to be connected. By definition, two nodes are neighbors if they are connected by a link. In other words, the clustering coefficient of a node is calculated by dividing the number of existing links between its neighbors by the number of possible links between them. Social networks are highly clustered. For instance, it is pretty common for people to have a friendship relationship with friends of their friends.

In 1998, Duncan Watts and Steven Strogatz proposed a new graph generating model [58]. The Watts-Strogatz model could produce networks with small average shortest path lengths which were also highly clustered. This model interpolates between regular lattices and fully random graphs produced by the Erdös-Rényi model. The model starts with a ring lattice containing n nodes, in which each node

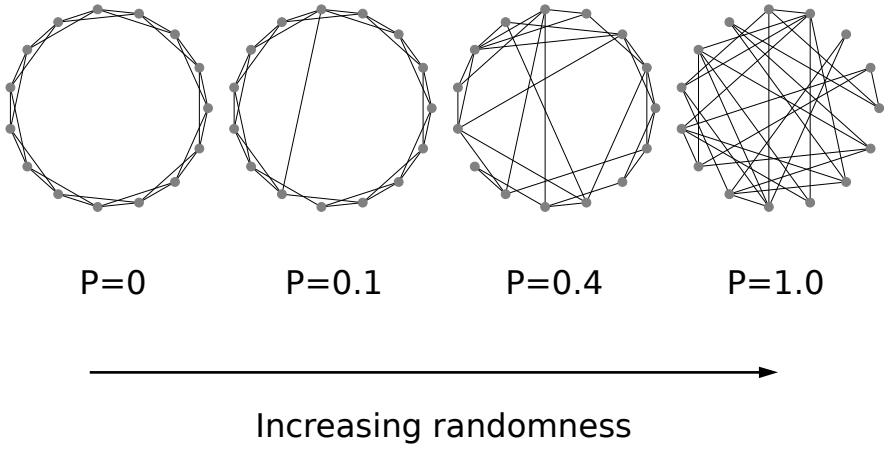


Figure 3: The picture shows how the Watts-Strogatz model interpolates between a regular ring lattice and a fully random graph.

is connected to k nearest nodes. Then a node is chosen and each link connected to it is rewired to a randomly chosen node with probability p and this step is repeated for all the nodes. The rewired graph has short path lengths because of the randomly rewired links and is highly clustered because of the regular lattice backbone (see Fig. 3).

2.5.3 The Barabási–Albert Model

As explained in the previous section, the Watts-Strogatz model produces graphs which are small-world: they have short path lengths and are highly clustered. However, many real-world networks have another property that the Watts-Strogatz model is unable to produce: they are scale-free. In other words, the *degree* distribution of nodes follows a power law, where degree is defined as the number of the neighbors of a node [55].

The World Wide Web (WWW) was the first network reported to be scale-free. In 1999, Réka Albert, Hawoong Jeong and Albert-László Barabási published a paper reporting observation of a "power-law tail" for the degree distribution of the WWW network [65]. The power-law degree distribution indicated the existence of a small number of hubs, nodes of high degree, while the majority of nodes of the network have only a few neighbors. Barabási, in the introduction of his 2016 book on network science [63], discloses that the result was shocking for them: they expected the degree distribution to be Poisson, not power-law, as it was the assumption in both the random graph and the social networks literature.

To examine if the scale-freeness is merely a characteristic of WWW networks or is

a universal feature, they repeated their study on three extremely different networks: the wiring diagram of a computer chip, a network of collaboration between Hollywood actors and a power grid network. To their excitement, the degree distribution of all these networks followed a power law. In 1999, they published a paper titled "Emergence of Scaling in Random Networks" [59] to report the observation of scale-free feature in a wide range of networks and also to propose a random graph generation model, known as the Barabási–Albert model (B-A model), to explain the phenomenon.

The B–A model relies on two ingredients, *growth* and *preferential attachment* [63]. In this model, the generation of graph starts with an initial seed, e.g. a clique (a fully connected graph) of a few nodes. Then in each step, m nodes are added to the network (growth) and each of these nodes connects to the existing nodes with a probability proportional to their degrees (preferential attachment).

The B–A model was neither the last nor the most realistic graph generation model. For example, empirical results have shown that unlike the networks produced by the B-A model, the probability density function of the nodes' degree in mobile call networks does not follow a power law with an exponent of -3 [11]. Moreover, based on the B–A model, when a network grows, hubs of unusually large degree become more probable. In practice, however, the number of social relationships that a person can handle is very limited because of time and cognitive limitations [25]. Furthermore, the B–A model assumes all the links to be identical, however, it is known that human relationships vary in intensity and have different functionality. Thus, a weighted network is more appropriate for modeling social networks.

Similarly to the model evolution procedure explained above, several models were developed after the Barabási–Albert to explain different aspects of real-world networks. What we should notice in here is the continuous feedback between real-world observation and mathematical models in network science [57]. The models and theories should be developed based on the real-world observations.

2.5.4 The Strength of Weak ties

In a social network, the links are not all identical but they have different features and functionalities. *The Strength of Weak Ties* [4], the paper authored by Mark S. Granovetter in 1973, brought the different functionality of *weak* and *strong* interpersonal ties under the spotlight. Granovetter, in this highly cited paper, describes the weak links as bridges which glue the otherwise isolated parts of the social network and as the channels through which socially distance information may reach the ego. As he reports, in the small-world experiment conducted by Milgram, the letters reached the recipients in fewer steps in those cases that they were passed to acquaintances instead of friends [4, 66].

Acknowledging that the social ties have different strengths, there have been several studies in the field of computational social science on various related topics such as how to measure the strength of ties [67], the topology of egocentric networks [13, 12], and also the topological structure of societal-level networks and how it impacts the dynamics of phenomena happening on the networks [36].

2.5.5 Egocentric Networks

An egocentric network consists of an individual (ego) and all his immediate neighbors (alters). Collecting data for forming an egocentric network is straightforward compared to collecting data for forming the social network of a group of people. For instance, in the case of using traditional data collection methods such as surveys, it is enough to interview one single person, the ego, to construct the network.

Trivially, the study of the structure of an egocentric network cannot reveal the structure of the entire society. However, for instance, sampling a large enough random set of egos, the average number of their alters would be a good estimation of the average degree in the entire social network [55].

Furthermore, the study of egocentric networks can give us insight on the behavioral traits of individuals. For instance, Ref. [12] investigates egocentric networks looking for persistent patterns indicating behavioral laws. In this study, the egocentric networks are constructed from outgoing call events of egos recorded in a CDR dataset. The links which are connecting each ego to his alters are weighted. The weight of each ego-alter link is defined as the total number of times the ego has called the alter during a time window of several months. The structure of weighted egocentric networks can be captured in the so-called *social signatures* [12]. Ref. [12] showed that each ego has a distinct social signature and these social signatures are persistent in time which means that despite large turnover in the membership of egocentric networks, their weighted structure remains more or less the same. The turnover in an ego's personal network is defined as the Jaccard index of the sets of the ego's alters in two consecutive periods of time (social signatures and the Jaccard index are elaborated upon Sec. 4).

2.5.6 Burstiness in Human Communication

The dynamics of social phenomena are governed by behavior of human individuals. Early models of human behavior assumed that the occurrence of the actions of individuals can be modeled as a Poisson process [68] and thus, the interevent times are exponentially distributed. This means that the times of occurrence of an individuals' actions are random. However, results of recent researches challenge this assumption and suggest that a wide range of human activities, such as communication [69, 70, 40], web-browsing [71] and creation of internet links [72], follow bursty patterns. This means that the time series of human activities consist of brief high-activity periods divided by long periods of inactivity.

The burstiness of human activity has been intensely studied recently [70, 69, 72, 73, 36]. These studies address varying aspects of this phenomenon such as the origin of burstiness in human behavior [70], different types of human activities with bursty characteristics [69, 71, 72, 74] and the impacts of burstiness on the well-known models of human activities [36]. Here, we review a few key studies on the mentioned topics.

In a 2005 article [70], Barabási reported that emailing activities of people follow a non-Poisson bursty pattern. He hypothesized that human communication is bursty simply because people do their tasks based on some perceived priority. Few months

after publication of that paper, another research work indicated a bursty pattern in lifelong letter correspondences of Charles Darwin (1809 -1882) and Albert Einstein (1879 - 1955) [69]. The research article showed that, despite the change in means of communication, the correspondence patterns of Einstein and Darwin are the same as the patterns observed in modern electronic exchanges like emails. It claimed that the reason is that Einstein and Darwin prioritized replying to their letters similarly to the way people nowadays prioritize answering their emails. The paper demonstrated that the interevent time distribution of letter correspondence is a power-law with an exponent of 3/2, while the power-law exponent for email correspondence was shown to be equal to 1 [70]. Based on these findings, the authors discussed that letter and email correspondences belong to different universality classes. However, a more recent paper [75] challenged the existence of these universality classes by showing that interevent-time distribution is a function of average activity and can change significantly when the activity level changes. In particular, it showed that in the case of letter correspondence of Einstein and Darwin, the calculated power-law exponents vary in different years and only the power-law exponent of the aggregated distribution is equal to 3/2. This result challenged the theory of the existence of universal classes claimed in [69].

The bursty patterns of human communications can have impacts on phenomena taking place on social networks. Social networks, as well as many other human-created or natural complex networks, show small-world properties. This means that the average shortest path between any randomly chosen pair of nodes is considerably small compared to the size of the network and most of the network nodes are located just a few links away from any randomly chosen node [76]. This property makes these networks topologically capable of rapid spreading. However, based on empirical data, spreading in social networks is surprisingly slow [77, 74]. In Ref. [73], the authors described how the bursty nature of the communication results in decelerating the spreading process of email worms. Considering the non-Poissonian interevent time distribution of sending emails, they predicted a decay time of 9 months for the prevalence of email worms. This result is in agreement with empirical observations [78].

Analogous to the diffusion of electronic viruses, epidemics and the spreading of biological viruses are also impacted by the dynamics of the communication behavior of human individuals. In Ref. [36], the authors used an empirical mobile-phone dataset to extract contact sequences and simulate the susceptible-infected model on the constructed network. They compared the result with a null model and suggested that bursty activity patterns of individuals slow down the spreading process significantly.

In order to explore the impact of human activity patterns on information diffusion, the authors of Ref. [79] conducted an email experiment. This experiment tracked the diffusion of a specific piece of information which was spreading by emails. The result of this experiment indicated that the diffusion of information is slow. However, this result cannot be predicted by rudimentary models which do not take the bursty behavior of humans into account.

2.5.7 Challenges in Detection of Social Ties and Measuring Their Strength

How to measure the strength of social ties? For decades this has been a topic of ongoing debates in the field of social network analysis. Granovetter, in his 1973 paper entitled *The Strength of Weak Ties*, hypothesized that the strength of interpersonal ties can be quantitatively defined as "a (probably linear) combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie" [4]. However, in that paper, Granovetter himself categorized ties to discrete categories of strong, weak, or absent and postponed the discussion on how to measure these four factors and how to calculate the strength by combining them to future studies. However, the future studies mostly brought more fundamental questions on measuring tie strengths instead of the practical method that Granovetter expected.

A challenge for defining tie strength is lack of reciprocity in social relationships. The two parties of a relationship can have very different perception of the strength of the tie between them. The difference can be to the extent that one of the two persons concerned questions the very existence of the tie [57].

In computational social science, it is a common practice to infer a social network from electronic records which contain information on communication events of a group of people. An example of these electronic records is call logs. When constructing social networks from call logs, the frequencies of calls are typically used as link weights and also as a proxy for tie strengths [11, 12, 13, 80]. In Ref. [12], in order to check if call frequency is an appropriate proxy for tie strength, the egos are asked to rate their *emotional closeness* to their alters. Then, the high correlation between the perceived closeness and frequency of calls is used to justify the consideration of call frequency as a proxy of tie strength.

However, there are several arguments implying that frequency of calls is not a *perfect* proxy for the strengths of ties. For instance Ref. [81] states that low frequencies of mobile-phone call and of SMS do not accurately identify weak ties. They report the existence of many alters that are perceived close by egos but are rarely called or texted. This might be a consequence of the fact that maintaining different types of ties needs different levels of communication effort. For example, Ref. [67] reports that old friendships tend to involve less frequent contacts between persons. Kin relationships are also reported to need less maintenance to stay at relatively high levels of emotional closeness [5, 81].

Moreover, even if we assume that communication frequency determines tie strength, calls are only one of the several channels that we use for communication and merely monitoring the call channel gives us an incomplete picture of people's communication behavior. Specially because as it is shown in this Thesis (Sec. 4.4.4), people use different communication channels to communicate with different sets of people. Therefore, the communication events via one single channel are not a representative sample of all communication events of a person. However, many times the available data is limited and contains the details of communication events via only one or two channels. Another practical challenge is combining the information on communication via different channels: the channels are intrinsically different, e.g.

phone calls have durations but text messages do not. This problem is addressed in this Thesis and a method is suggested for combining the information on communication via different channels in order to construct a single weighted network from them (Sec. 4.1).

Moreover, when aggregating communication events over time–like counting the number of calls between each pair of people during a time window–the length of aggregation window also impacts the characteristics of the resulting network. For instance, a short time window cannot capture a considerable portion of ties. Especially since the human communication is bursty, the interevent times can be long. As a result, by increasing the aggregation window the network grows slower than the hypothetical case with Poissonian interevent times. Ref. [82] investigates the impact of the length of time window on the aggregated call network. The article reports that by increasing the aggregation window from the initial length of one day, first mainly clusters of strong links appear and then gradually the network grows and weaker links also appear.

To conclude, when using electronic records to infer concepts such as tie strength, it is vital to keep in mind what we are measuring and how it is only a proxy of the concept we want to study.

3 Datasets

The number of mobile phone users has increased dramatically since the beginning of the 21st century. In 2014 the number of mobile phone subscriptions around the world reached 96% of the world population as compared to 12% in the year 2000 [34]. Moreover, mobile phones are used as personal devices (they are usually used by one person only) which people carry around almost everywhere. As a result, they can serve as sensors to capture the behavior of human individuals.

Call detail records (CDRs) datasets are a basic but rich source of data on communication behavior of humans. CDRs are collected by mobile phone operators for billing purposes and they contain detailed information on the communication events of the customers. This information includes the phone number or hashed ID of the two participants of communication, the type of the event (e.g. phone call or text message), the timestamp of the event, the duration of call or the length of text message, and possibly the location of the cell tower which the user who initiated the communication was connected to. Additionally, some basic information about users like age and gender may be available in conjunction with CDR datasets.

The main advantage of CDR datasets is that they are usually enormous in size and can contain a big part of the population of a whole country. For instance, the largest dataset studied in this Thesis, which is also a CDR dataset, contains communication data of 9,674,264 subscribers.

However, the golden age of CDR datasets can be considered to be over. Smartphones, which are getting constantly more and more popular, have provided mobile phone users the possibility to use various communication channels and mobile phone applications for socializing. As a result, a smaller share of people's communication takes place via the call and text message channels. Thus, CDR datasets are less representative of the social life of users than they used to. This has motived researchers to collect more comprehensive mobile phone datasets using mobile phone applications designed for data collection purposes (see, e.g., Refs. [50] and [52]). These types of datasets are limited in population size compared to CDR datasets but are higher in resolution and contain more details.

In the following sections, in Sec. 3.1 and 3.2, we describe the characteristics of the two CDR datasets used in this Thesis.

3.1 The large mobile phone dataset (DS1)

The largest dataset used in this work is a call detail record (CDR) dataset gathered by a telecommunication company in a European country for billing purposes (see, e.g., Refs. [36, 38]). This dataset contains data on call and text messages of 9,674,264 anonymized subscribers. We will refer to this dataset as DS1 for convenience. The data we received consists of a few subsets whose their collection dates range from 2007 to 2009. In this work, I have used a subset which contains communication data from the beginning of January 2007 to the end of July 2007. I chose this subset because unlike other subsets in DS1, it also contains information on communication events between company and non-company users. The company users are those

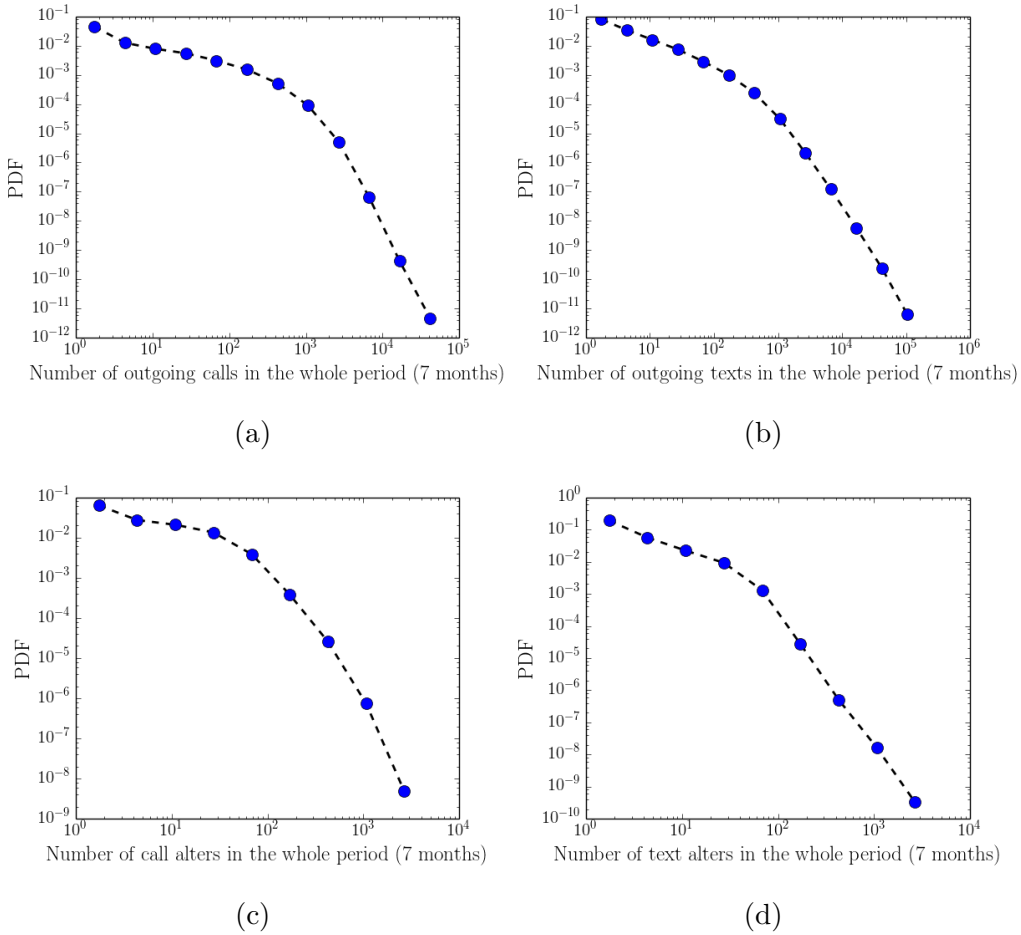


Figure 4: The probability density function of total number of phone calls (panel (a)), total number of text messages (panel (b)), the number of called alters (panel (c)) and the number of texted alters (panel (d)) during the 7 months period in DS1.

users who were subscribed to the operator, so that we have all information on their outgoing communication events. However, non-company users are present in the dataset only if they have been contacted by company users.

The dataset DS1 consists of two categories of files: first, the events files which include information on outgoing communication of company users and second, the contract files from different years which associate some demographic and geographic information to the user IDs. Moreover, there is information available on if each user ID represents a unique individual contract or a group family contract.

There are 5,846,643 users in the dataset after the basic preprocessing (described in Sec. 3.1.1). The distributions of the number of calls and texts and the number of alters across the population are shown in Fig. 4.

3.1.1 Basic Preprocessing

We use CDR datasets as a source to study and analyze the behavior of human individuals. Thus it is desired that each user ID represents a unique individual during the whole data collection period. However, many times this is not the case: each hashed user ID represents a phone number which can be used by different individuals at the same time, for example family contracts, or owned by different individuals during different periods. We tried to remove these cases when preprocessing the data. We only included those IDs which had unique individual contracts and also excluded hashed IDs related to phone numbers which based on demographic information (age) had changed ownership between 2007 and 2009. The logic behind the latter filtering is that those phone numbers which have been used by the same individual (at least based on the consecutive contract files) are less probable to have changed ownership during the 7-month period that we are studying.

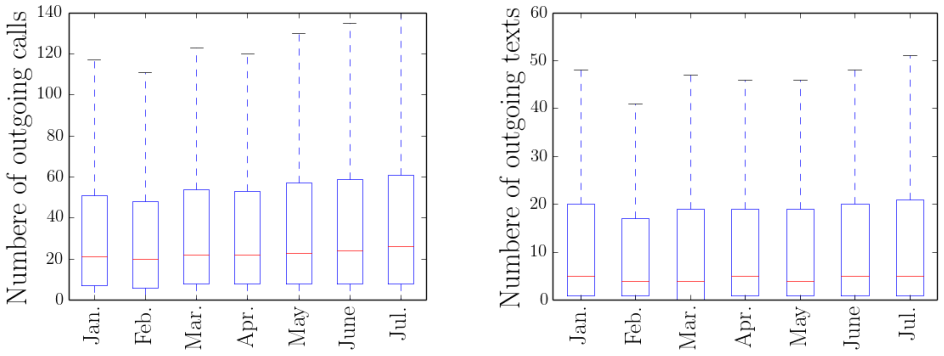
Moreover, we filtered out all the call events which had negative or non-integer values as their durations which counted less than 1% of events (in total 13,590,186 events). The total number of users in the dataset after the basic preprocessing is 5,835,067.

3.1.2 Filtering Based on Activity Level

In this Thesis, we have been studying patterns of social behavior of individuals in time and across call and text communication channels. Hence, we need to pick the users who have been actively using the communication channels throughout the period of study. We did this by setting a monthly activity threshold of 20 phone calls and 7 text messages. The number of users in the dataset after applying these thresholds reduces to 506,331 (around 12% of users). The activity thresholds (20 calls and 7 text messages per month) are approximately equal to the median of the monthly number of outgoing phone calls and text messages across the population (see Fig. 5).

	DS1
Number of active users	506,331
Length of data-collection period	7 months
10-percentile of NCPM*	43
Median of NCPM	83
90-percentile of NCPM	194
10-percentile of NTPM**	20
Median of NTPM	45
90-percentile of NTPM	131

Table 1: The summary statistics on calling and texting activities among the active users in DS1. NCPM stands for number of calls per month and NTPM stands for number of text messages per month.

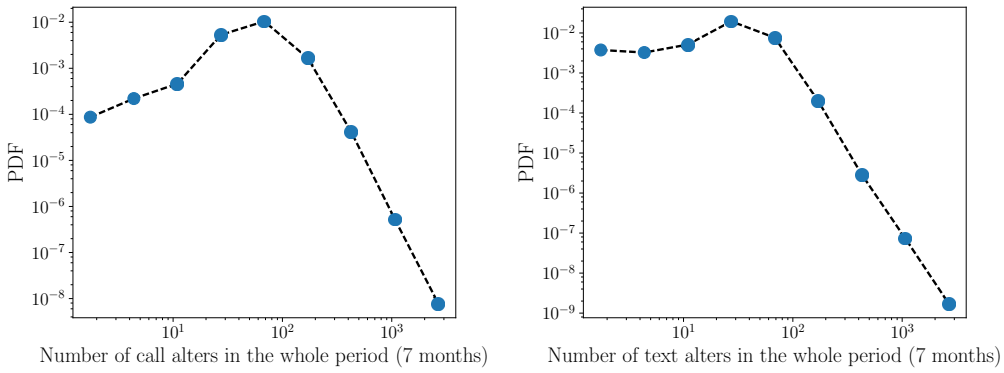


(a) (b)

Figure 5: Distributions of the number of outgoing calls in DS1 for each month among the users (panel (a)) and the distribution of the number of sent text messages for each month among DS1 users (panel (b)).

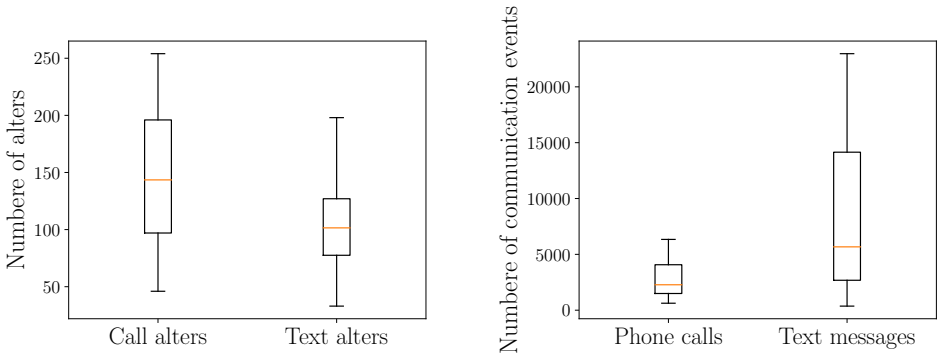
3.1.3 Data Statistics Among Active Users

In the previous section, we explained the basic preprocessing done to clean the data and the filtering based on activity levels to pick a suitable subset of the users for the purposes of our studies. The total number of communication events in the remaining data is 613,744,524. A summary of basic statistics of the data after the filtering can be found in Table 1. Also, the probability density function of the number of call and text alters among the active users can be found in Fig. 6.



(a) (b)

Figure 6: The probability density function of total number of called alters (panel (a)) and total number of texted alters (panel (b)) during the 7 months time-period among active users in DS1.



(a) (b)

Figure 7: Boxplots of the total number of called and texted alters during the 18-month period in DS2 (panel (a)) and boxplots of the total number of communication events during the 18-month period (panel (b)) among the users in DS2.

3.2 The UK Students Dataset (DS2)

The other dataset used here is a small dataset on calls and texts of 24 students over 18 months [6]. The dataset is collected in the UK starting in 2007 when the subjects were in their last year of high-school. This was a period of change and of high turnover in the social networks of students: some of them moved to another city, started university studies or got a job. Because of this, the dataset is ideal for studying the impacts of social changes on the structure of egocentric networks.

The communication information in this dataset has been extracted from the participants monthly phone invoices by their permission (containing the types of the communication events, the timestamps of the events, the phone numbers of the recipients and the duration in case of a call event). Moreover, the dataset is extended by questionnaires which have been filled by students every 6 months. The questionnaires that the students filled asked some questions about their alters such as the types of their relationships and their emotional closeness to each alter; in this Thesis, we have not used the questionnaire data. The initial number of participants in the experiment was 30 (15 females, 15 males). However, only 24 of them (12 females, 12 males) have used their phones throughout the study and filled all the questionnaires.

Compared to DS1, this dataset is smaller in size and contains communication events of only 24 individuals, which is in total 369,919 communication events. On the other hand, compared to DS1, DS2 is longer in time and all the 24 subjects are actively using their mobile phones. A summary of statistics on calling and texting of the subjects in DS2 can be found in Table 2. Also, the distributions of the total number of alters and the total number of communication events among the subjects are shown in Fig. 7.

Table 2: The summary statistics on calling and texting activity among the active users in DS2. NCPM stands for number of calls per month and NTPM stands for number of text messages per month.

	DS 2
Number of active users	24
Length of data-collection period	18 months
10-percentile of NCPM*	67
Median of NCPM	127
90-percentile of NCPM	278
10-percentile of NTPM**	105
Median of NTPM	317
90-percentile of NTPM	2019

4 Methods and Results

Humans are social animals. In the course of evolution, our ancestors have formed purposeful groups such as families and tribes and invented languages to communicate [83]. Anthropological research has shown that proto-humans had larger brains compared to other primates, which enabled them to form larger groups [84, 85]. Then language emerged to make cooperation in the large groups possible and to allow the members to share information and ideas. Later in history, modern humans developed social norms and rituals and established complex social structures such as villages and cities.

Nowadays, in the modern age of communication, we have developed technologies which enable us to communicate, to socialize, and to get information fast and easily. The widespread use of digital technologies such as mobile phones and the World Wide Web, not only has made the humanity interconnected but has also resulted in the production of enormous data on human behavior. The availability of digital traces on human behavior brings the opportunity to study human beings both on the individual and the societal levels [34, 27, 24].

In this Thesis, we have used two mobile phone communication datasets (Sec. 3) to study the social patterns of people. These datasets consist of information on time-stamped communication events. Given time-stamped data on communication between individuals, we can construct a network where each node represents a human individual and each link can indicates communication or a type of relationship between a pair of individuals. To do so, we should first answer this critical question of how to define our network. Depending on what we want to study, differently defined networks are useful for us. For instance, temporal networks are strong tools for studying the dynamics of human interactions, such as the propagation of information or diseases [86, 87]. On the other hand, a statical aggregated network, made from information on social events over a period of time, can provide us insight on the overall social structure [11]. For instance, it is shown in Ref. [12] that number of phone calls a person has made to each of her contacts during a period of time can be used as a proxy of her perceived emotional closeness to the contacts.

In this Thesis, we use aggregated networks because our focus is on social relationships (ties) instead of individual social interaction events that take place on fast time scales. In particular, we focus on egocentric networks, that is, one focal node (ego) and the ego's network neighbors (alters).

4.1 Defining Egocentric Networks

People form and maintain their relationships through a diversity of communication channels, for example, face-to-face interactions, phone calls, text messages, emails, and interaction on social media. Since these channels are different in nature and functionality, individuals do not use them interchangeably. Many factors contribute to an individual's choice of communication channel. For instance, the type of the relationship (nature of the social tie), the general channel preference of the individual (characteristic of the node), the time of the event (social norms) and the reason

for communicating. As a result, when constructing egocentric networks, the more channels we include the more realistic picture we get. However, the available data are often limited and do not include information on communication events through all the channels. Moreover, even when information on communication through several channels is available, combining them is problematic. The reason is the intrinsic differences between the channels. For instance, the duration or frequency of calls to each alter is typically used as a proxy for the tie strength (the link weight) [27, 14, 11, 80, 13]. But text messages have no duration, and the number of text messages between an ego-alter pair is not directly comparable to the number of calls between that pair. While a conversation can take place during one single phone call, usually several ping-pong text messages are exchanged during a conversation [42].

In this Thesis, we have developed a method which enables us to make channels more comparable. This method uses the timestamps of the communication events and by coarse-graining the timelines allows us to define call-based ego-networks and text-based ego-networks which are comparable and also make combined ego-networks of information on both of the channels. The steps are as follows (see Fig. 8):

1. Divide the communication timeline of each ego-alter pair to time-bins of one hour.

2. Calculate the link weights:

   - To define link weights in the text message ego-network, count the number of one-hour time-bins which contain at least one text message for each ego-alter pair. Thus, for example, a link weight of $w = 8$ indicates that there were 8 one-hour time bins in which texting activity has occurred.
   - To construct the call ego-network, count the number of time bins containing at least one phone call. To take the duration of phone calls into account, if a phone call is stretched over several time bins, they should all be counted.
   - To define the link weights in the combined ego network, count the number of time bins during which at least one communication event of any type took place (either text message or phone call).

An advantage of this method is that it can be used to calculate link weights that quantify the amount of communication or social interaction in any channel, as long as the time stamps of communication events are available.

4.2 Social Signatures

Maintaining social relationships is costly and we have finite resources for socializing: both time and our brain capacity are limited [13, 10]. We divide these limited resources unevenly among our alters: a few strong ties receive a large proportion of our resources while the remaining are divided among a large number of weak ties. This disparity is reflected in the topological structure of egocentric networks as well as in the so-called *social signatures* [12].

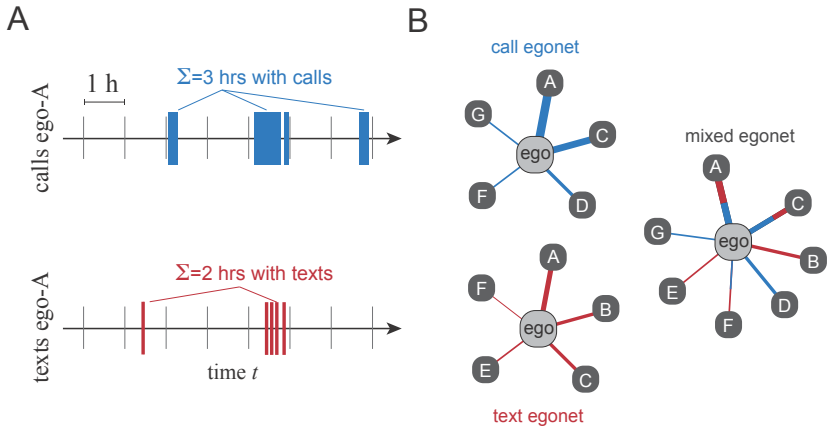


Figure 8: Building egocentric networks from phone call and text message communication records using time-binned weights. Panel A: The timelines corresponding to each of the ego-alter relationships are divided into short time-bins (for example of one hour). Then, the number of bins with at least one communication event of the desired type(s) (call, text, call/text) is counted. These numbers determine the link weights in egocentric networks (panel B). The figure is taken from Ref. [88].

The social signature of an ego measures the fraction of communication dedicated to alters when they are ranked according to this fraction. Given the weighted ego network of an individual, the social signature is constructed as follows: we rank the ties based on the link weights so that the first rank is associated with the highest weight. Then we normalize the weights by dividing them by the sum of all link weights (see Fig. 9). The social signature of the ego i then reads:

$$\sigma_i = \{(w_{i1}/\sum_{j=1}^{k_i} w_{ij}), \ldots, (w_{ik_i}/\sum_{j=1}^{k_i} w_{ij})\}, \tag{1}$$

where the alters j are sorted by link weight in decreasing order and k_i is the degree (total number of alters) of the ego i.

It was shown in Ref. [12] that each individual has her own, distinctive social signature that persists in time, even when there is a large turnover in the ego's network. This is the reason why these patterns are called *signatures*. The social signatures analyzed in Ref. [12] are made based on the number of phone calls to each alter in the Students dataset (see Sec. 3.2). Ref. [14] also has reported similar results on another dataset of mobile telephone calls. Another work has studied social signatures made from email egocentric networks and observed that they also are persistent [15].

In this Thesis, we will test the persistence of phone call, text message and mixed social signatures. Moreover, we will study differences and similarities of these three types of social signatures.

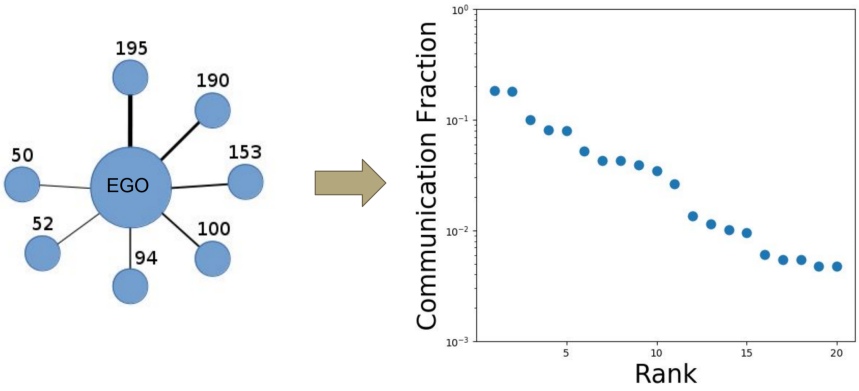


Figure 9: People do not distribute their communication effort uniformly among their ego-network members: a few closest alters get a disproportionately large fraction. This results in social signatures that typically decay slower than exponentially (the right panel). Figure after Ref. [12].

4.3 Comparing Social Signatures

In the previous section, we described social signatures as distinct and persistent patterns of individuals' social behavior. To be able to talk about persistence and similarity, we should first define a distance function to measure the distance between a pair of signatures, which are sorted arrays of fractions which can vary in length: different egos can have different numbers of alters.

Based on the mathematical definition, a distance function defined for a set (here the set of social signatures) maps each pair of elements of the set to a non-negative real number. So

$$d : S \times S \to [0, \infty) \tag{2}$$

where d stands for distance and S is the set of social signatures. The following conditions should be satisfied: d should be non-negative; the distance between two elements should be equal to zero if and only if they are identical; the distance function should be symmetric: $d(s_1, s_2) = d(s_2, s_1)$, and the triangle inequality should be fulfilled: $d(s_1, s_2) \leq d(s_1, s_3) + d(s_3, s_2)$.

In Secs. 4.3.1 and 4.3.2, we introduce two distance functions used in the literature to measure the distances between pairs of social signatures. In this Thesis, we use the Jensen-Shannon distance function which is explained in Sec. 4.3.2.

4.3.1 L^2 Distance

L^2 or Euclidean distance is a basic distance function. When we talk about distance in everyday life, we usually refer to the two or three dimensional L^2 distance. In order to use L^2 distance to measure the distance between a pair of social signatures, s_1 and s_2, first if the social signatures are not of the same length, we append zeros to the shorter array so they become of the same size. Then the L^2 distance between

s_1 and s_2 is defined as:

$$L^2(s_1, s_2) = \sqrt{\sum_{r=1}^{k} |f_{1r} - f_{2r}|^2}, \tag{3}$$

where f_{1r} is the fraction of communication that the alter of rank r in the signature s_1 receives.

4.3.2 Jensen-Shannon Distance

The Jensen-Shannon divergence (JSD) is a method for measuring how a pair of probability distributions diverge from each other. JSD is a smoothed and generalized version of the Kullback-Leibler divergence (KLD), so it can handle zero probabilities. Another advantage of JSD over KLD is that it is symmetric. This means that the square root of JSD can be used as a distance function [89].

The square root of JSD which is the Jensen-Shannon distance ($JS_{distance}$), has been used in Ref. [12] as a distance function to measure the similarity of pairs of social signatures. Similarly to the process of measuring distances with the L^2 norm, we first append a zero-pad to the shorter social signature, so that both of the signatures are of the same length. Then the $JS_{distance}$ between social signatures s_1 and s_2 is defined as:

$$JS_{distance}(s_1, s_2) = H(\frac{1}{2}s_1 + \frac{1}{2}s_2) - \frac{1}{2}[H(s_1) + H(s_2)], \tag{4}$$

where $H(s_1)$, the Shannon entropy of s_1, is defined as:

$$H(s_1) = -\sum_{r=1}^{k} f_{1r} \log f_{1r}, \tag{5}$$

where k is the maximum rank and f_{1r} is the fraction of communication dedicated to alter of rank r.

The $JS_{distance}$ between a pair of social signatures is maximized if one of them is a uni-friend signature with all the communication directed to one single alter and the other signature is flat, which means that all the alters are getting the same share of communication. The value of the maximum distance is a function of the length of the flat signature (see Fig. 10).

4.3.3 Jaccard Index

In Secs.4.3.1 and 4.3.2 we explained two distance functions which can be used to measure the difference between the *shapes* of two given social signatures. To compare egocentric networks with respect to their *membership* and to measure similarity of their sets of alters, we can use the Jaccard index. The Jaccard index or the Jaccard similarity coefficient is a measure of similarity between two finite sets. The Jaccard index between two sets A and B is defined as

$$\frac{|A \cap B|}{|A \cup B|}, \tag{6}$$

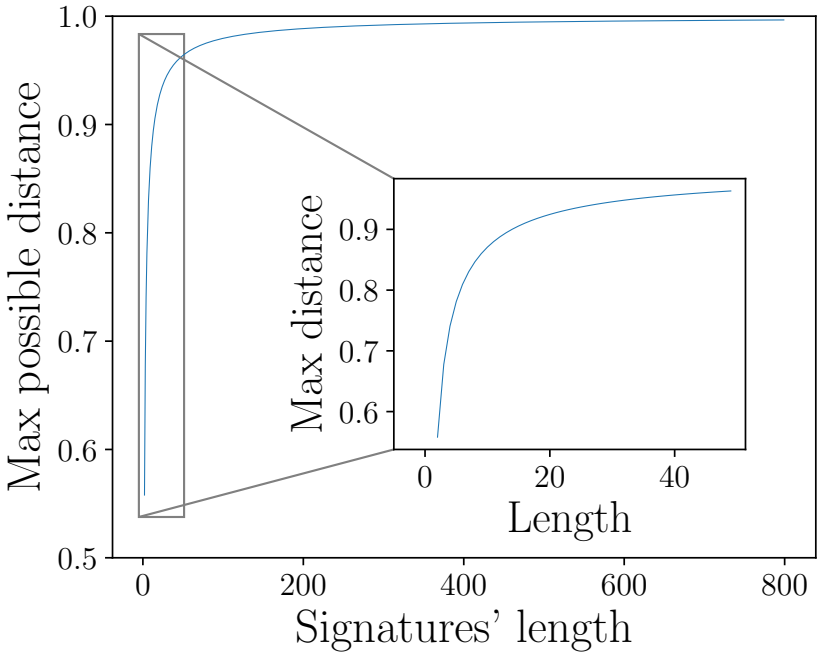


Figure 10: The $JS_{distance}$ between two social signatures is maximized when one of them is flat and the other one a uni-friend signature. The maximum distance is a function of length of the flat signature and rapidly approaches to one when length of signatures increase.

where $|A \cap B|$ is size of the intersection set of A and B and $|A \cup B|$ is size of the union set of A and B.

The Jaccard index is minimized (equal to zero) when the two sets have no members in common and is maximized (equal to one) when the two sets are identical. Thus, for example, the low value of Jaccard index between social signatures of an individual from two consecutive periods indicates that there is high turnover in her egocentric network.

4.4 Results

4.4.1 Call, Text and Mixed Social Signatures Are Persistent in Time

Each individual has a distinct, persistent social signature. This was shown in Ref. [12, 14] for number-of-calls social signatures and in Ref. [15] for signatures calculated from email communication. In here, using the same method used in Ref. [12], we test if the social signatures computed from call, text and combined egocentric networks defined in Sec. 4.1 have the same property. The steps are as follows:

1. First, we divide each dataset into two equal-sized time windows based on timestamps of the events: DS1 to two consecutive 3.5 months time windows

and DS2 to two consecutive 12-month time windows.

2. Then in each time window, for each ego-alter relationship, we divide the timeline into time-bins of one hour. Then we count the number of time-bins with at least one communication event of our desired type (call, text, call or text) to define link weights in egocentric networks (check Sec. 4.1. to see how the ego-networks are constructed in details.)

3. From call, text and combined ego-networks, we construct accordingly call, text and mixed social signatures (see Sec. 4.2). Now we have $2 \times 3 \times n$ signatures; n is the number of egos, multiplication by 2 is because of the number of time windows and multiplication by 3 is because we are making three different types of signatures.

4. Finally, we check if the social signatures are distinct and persistent. For example to check this for call signatures, we measure the distance between the call signatures of each individual in the two consecutive time windows (we use the JS-distance, see Sec. 4.3.2). We refer to these distances as *self-distances*. The distribution of self-distances among the population gives us a picture of how people's social signatures change in the time. Then to have a reference, we calculate distances between call signatures of different egos. We refer to these distances as *reference distances*. We plot the self-distance and between-distances on top of each other to compare the two distributions. Fig. 11 shows that for all three types of signatures in both of the datasets, the self-distance and reference distance distributions are significantly different.

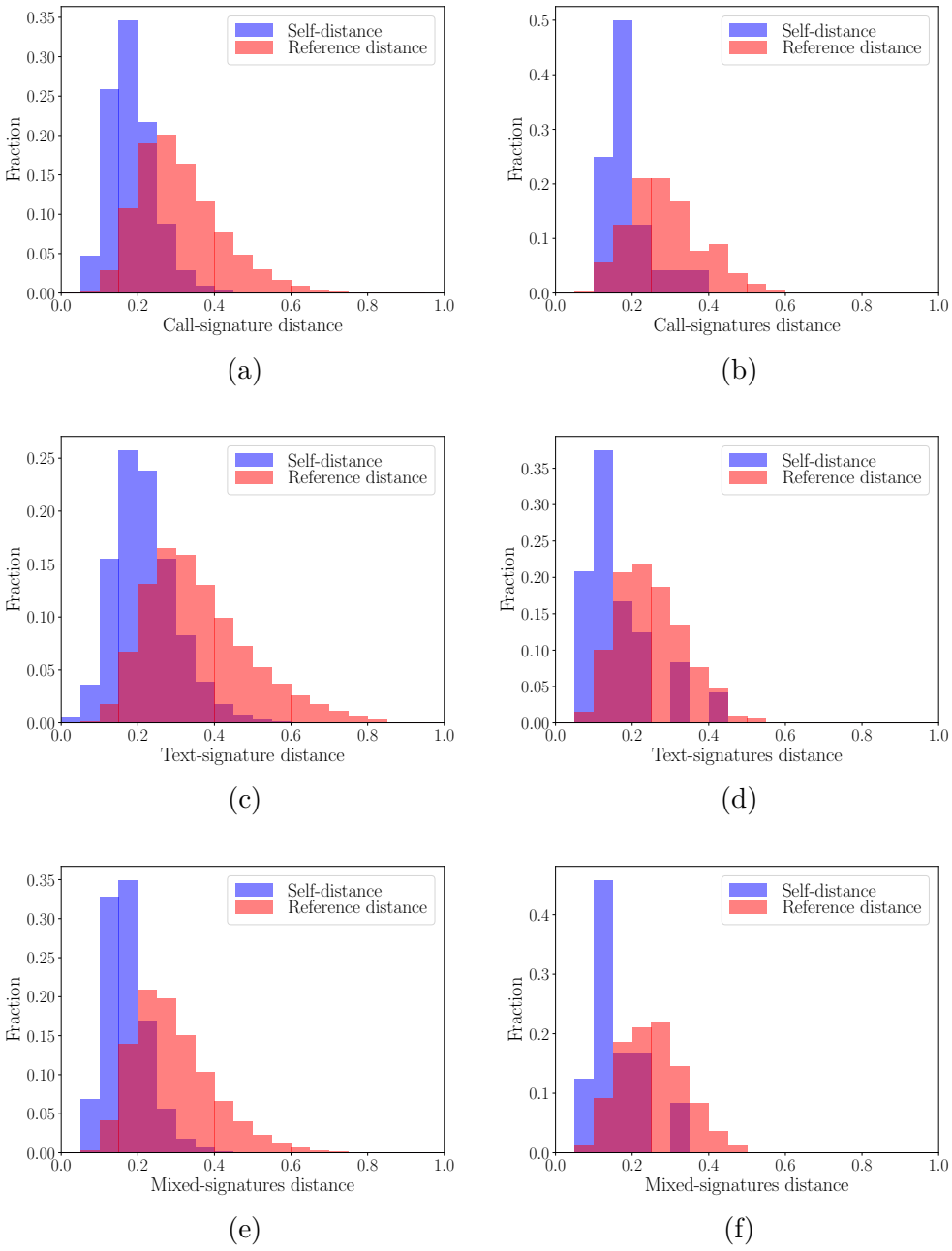


Figure 11: Shapes of social signatures persist in time at the individual level. This is true for both call and text channels as well as mixed signatures which are built based on both of the channels. Panels (a), (c), (e) show signature persistence in DS1 and panels (b), (d), (f) show the same phenomenon in DS2. The blue distributions illustrate distances between social signatures of each ego in two consecutive equal-sized time-windows (self-distances). The distributions of distances between social signatures of different egos which are used as reference distributions are shown in red (reference distances). Comparison of distributions of self-distances with reference distances verifies that call, text, and mixed signatures are persistent, because the self-distances are on average smaller than the reference distances. The plots are taken from Ref. [88].

4.4.2 Call and Text Signatures Have Similar Shapes, Both at the Population and at the Ego Levels

In the last section, we showed that people have persistent call, text and mixed social signatures. In this section, we compare these three types of social signatures at both the population and the individual levels.

First of all, we observed that the inequality in the division of communication effort among alters is reflected in all three types of social signatures. This can be seen in Fig. 12a which shows a call, text and mixed signatures all belonging to one example person as well as the population-averaged social signatures in DS2 which are shown in 12b.

Moreover, by looking at the two plots 12a and 12b, it seems that call, text and mixed social signatures are fairly similar in shape. This raises the question if we distribute our communication effort among our text alters similarly to how we distribute our phone calls among our call alters.

To check this, by using Jensen-Shannon distance function, we calculate the distances between call and text signatures of each ego (self-distances) and compare the distribution of self-distances with the distribution of distances between the call and text signatures of different individuals as a reference. For this comparison, we used social signatures made from aggregated ego-networks over the entire data collection periods (7 months in the DS1 and 18 months in the DS2). The comparison of resulting distance distributions confirms that self-distances are on average smaller than reference distances. This holds for the results on both of the datasets (see Figs. 12c and 12d).

4.4.3 Call and Text Egocentric Networks Differ in Composition

In the last section, we observed that call and text social signatures of an ego are similar in shape. However, in this section, we show that despite of the similarity in shape of signatures, the ego-centric networks are noticeably different in composition. This means that even if egos do not call and text the same sets of alters, both types of signatures have similar shapes: if only a few people get most calls in someone's personal network, then only a few people get most texts, even if those are not the same people, and if the call network is more flat, then the text network is more flat too.

The differences in the composition of call and text egocentric networks are made clear by the low values of Jaccard indices between membership sets (see Figs. 13a and 13b) This means that there are a lot of people that we only call or only text. To check if these results are not just a consequence of the long tails of the social signatures – all those alters who we have communicated only once or twice – we also measured the Jaccard indices between the top 20 alters in call and text ego-networks (see Figs. 13a and 13b). The values of the Jaccard indices were still low, which indicates that the core sets of alters in the two communication channels are typically significantly different.

We also observed that the ranks of those alters who are a member both call and text ego-networks correlate only moderately: an alter who is among the top alters in

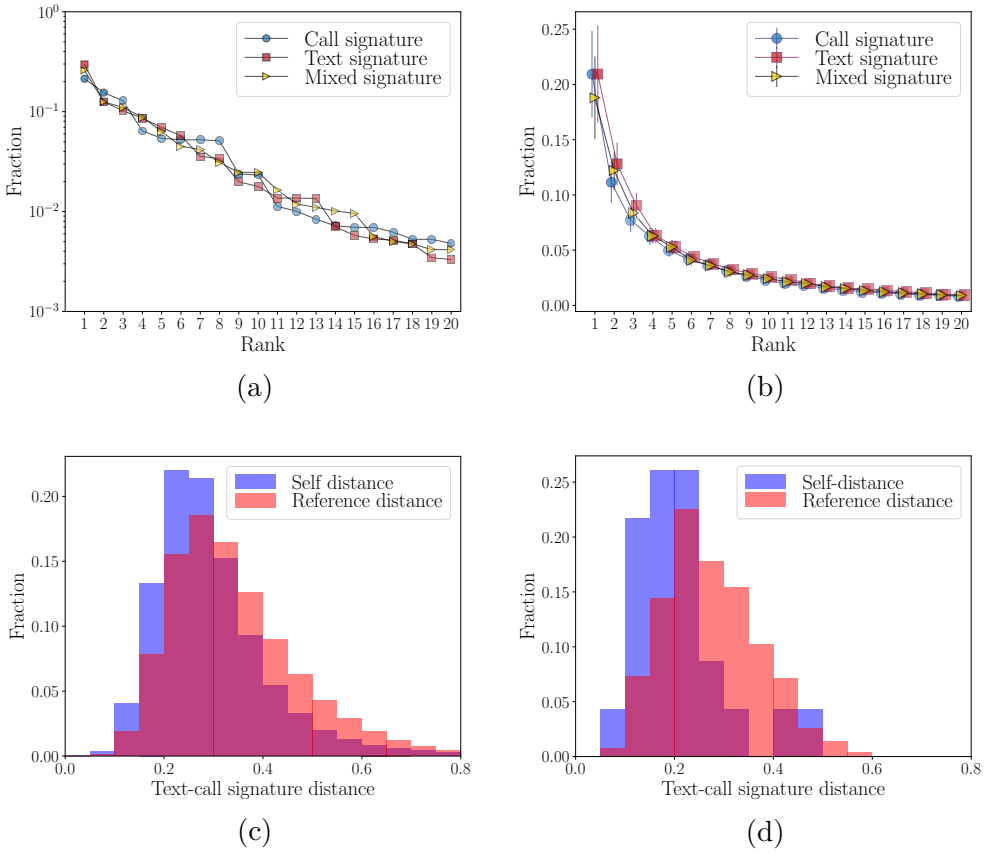


Figure 12: The similarities of social signatures of different types at the population and individual levels. Panel (a) illustrates the call, text and mixed signatures of one example person in the DS1. The three signatures are rather similar in shape. Panel (b) shows the averaged social signatures over the population in DS2. The population-level signatures also look similar. Panels (c) and (d) compare the distance distributions of the call and text signatures of same egos (in blue) with the distance distributions of call and text signatures of different individuals as a reference (in red). On average, the call and text signatures of each ego are more alike than pairs of signatures of different people. The plots are taken from Ref. [88].

the call signature may receive a far smaller share of text messages and vice versa (see Figs 13c and 13d).

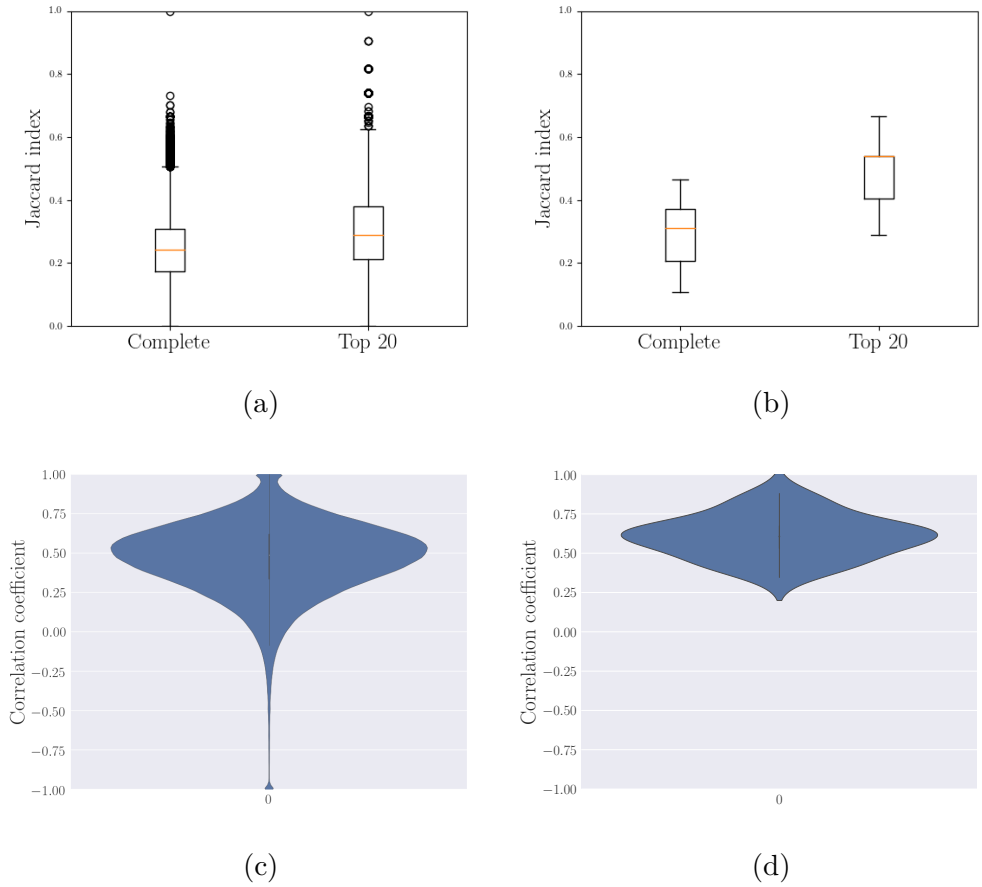


Figure 13: Although egos have relatively similarly-shaped call and text signatures, the egocentric networks formed via these two channels have low membership overlaps and the rankings of the mutual members are not the same. The distribution of Jaccard indices between the sets of top 20 call and text alters, as well as between all call and text alters, are shown in panel (a) for DS1 and panel (b) for DS2. The distribution of correlation coefficients between text ranks and call ranks of those alters who are a member of both of the ego-networks is shown in panel (c) for the DS1 and in panel (d) for the DS2. The plots are taken from Ref. [88].

4.4.4 Composition of Mixed Social Signatures: Channel Choice Does Not Depend on Alter Rank

To find an explanation for the similarity of call and text social signatures, we take a closer look at the composition of mixed social signatures. For instance, if the contribution of calls to the total communication share of the alter is the same for all the alters (for example, if all the ego's alters are 20% of time communicated by text and 80% of time by call), the call and text signatures would consequently be identical. However, we observe that the choice of channel is neither regular nor predictable from the alters' rank.

Fig. 14 shows an example mixed signature and its weight composition. We should notice that since the link weights are calculated by counting time-bins, there are overlaps between call and text shares: there are time-bins in which the same alter is both called and texted. The composition of this example mixed signature shows no clear pattern. It seems that the choice of the communication channel is not dependent on the alter's rank, but it is determined by some specific characteristics of each social tie.

This is confirmed by Fig. 15 that illustrates the shares of texts in all ego-alter relationships of DS2 (top) and DS1 (bottom). The only systematic feature seems to be that alters at top ranks are more probable to be contacted by both of the channels and the fraction of call-only and text-only ties increases at the lower ranks and towards the tails of mixed signatures. Beyond this, there are no systematic trends dependent on alter rank.

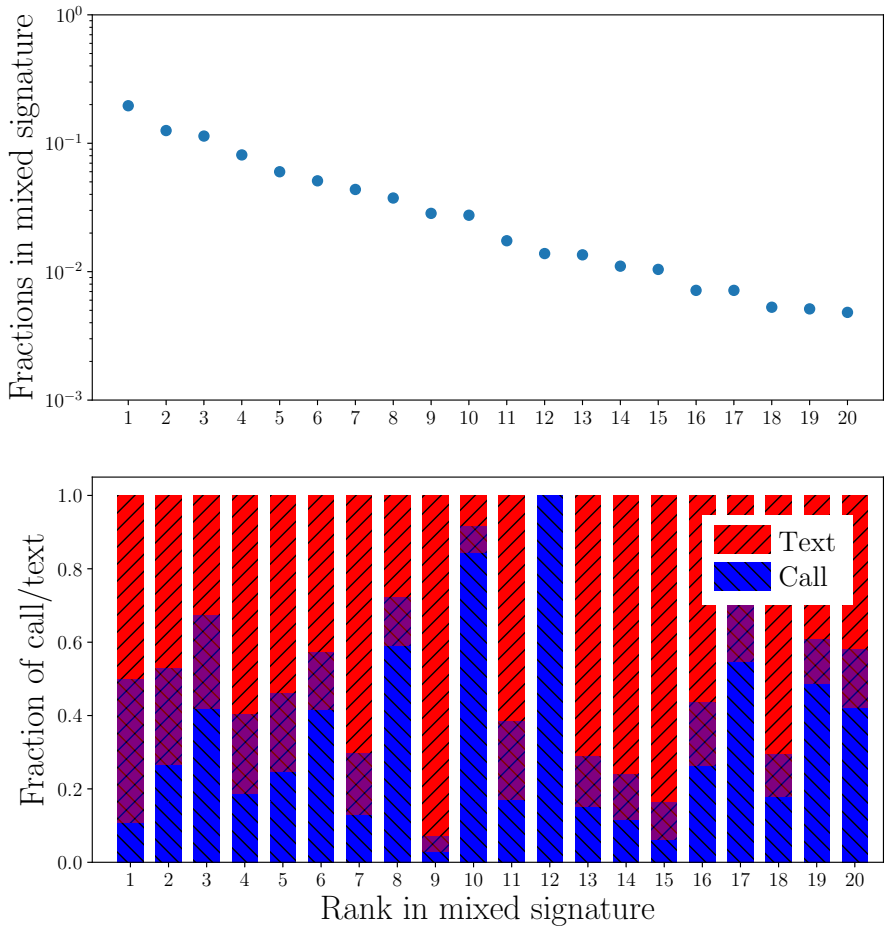
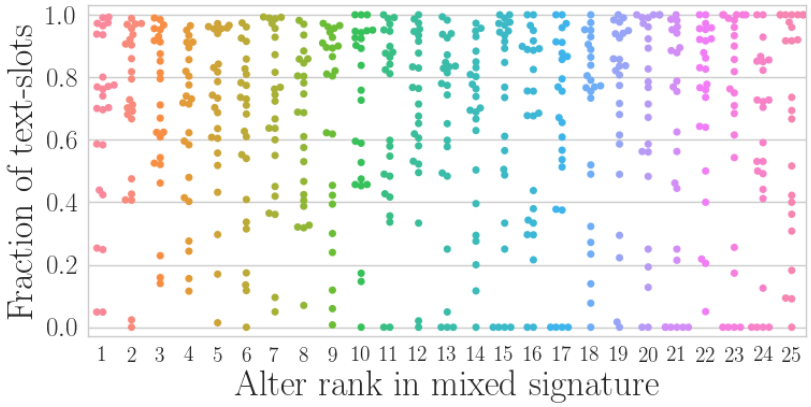
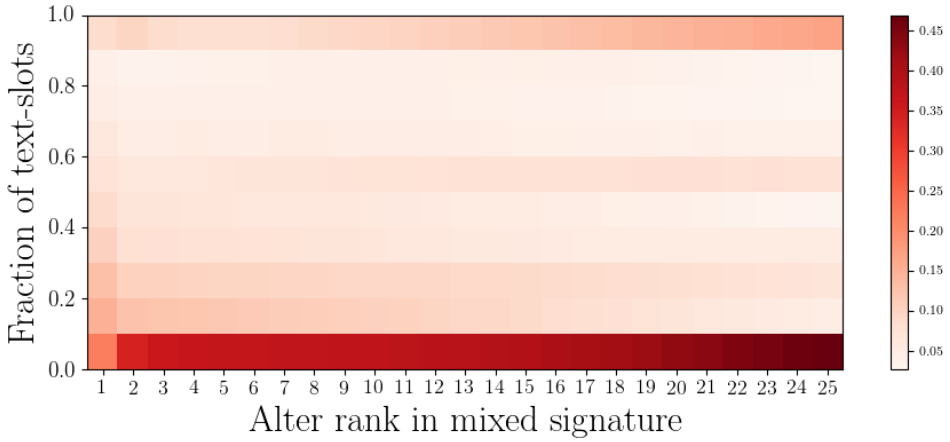


Figure 14: Top panel: the mixed signature of one of the egos in DS2. Bottom panel: the composition of the same mixed signature. The blue and red areas of the bars represent, respectively, the fractions of one-hour time-bins containing at least a text or a call. The purple areas represent the fraction of time-bins with both calls and texts. The figure is taken from Ref. [88].



(a)



(b)

Figure 15: Panel (a): Each dot shows, as a function of rank, the share of text channel in the communication between an ego-alter pair, or more precisely the fraction of one-hour time-bins containing text communication divided by the number of time-bins containing communication of any type. The plot contains the fractions for all ego-alter pairs. No general pattern can be seen, except that the alters of the top ranks are more probable to be contacted by both phone calls and text messages. On the other hand, in the tails of the signature (lower ranks), there are more alters who are merely called or texted. Panel (b): A heat-map version of the top panel for DS1, with intensity of colors indicating the number of ego-alter ties with a given fraction of texts at each rank. In DS1, compared to DS2 which is a sample of teenagers in age 17-19, text messages are used much less. The plots are taken from Ref. [88].

5 Summary and Discussion

Technological advances in digital computing and communication technology have brought humans to the information age. We are surrounded by digital devices and are becoming increasingly dependent on them. When using digital devices, as a byproduct, we produce an enormous amount of data. These detailed behavioral data fuel the recently emerged field of Computational Social Science [24].

Studying the interactions between people has been a topic of interest in both the traditional social sciences and computational social science. Passively collected electronic data such as mobile-phone datasets or email logs have been used to study the properties and the dynamics of social networks. These analyses are often done using networks constructed from data on communication via a single channel–which provides us with an incomplete picture. Such studies are many times limited by the availability of data. However, even when the information on communication via several channels is available, combining data from different channels to infer a single network is challenging. The challenge is caused by the intrinsic differences of the communication channels. In this Thesis, we suggested a method which enables us to compare and combine communication across different channels. As long as the timestamps of the communication events are available, this method can be applied to any communication channel, e.g., mobile phone calls, text messages, communication on social media, or face-to-face interactions captured by devices such as Sociometric Badges [90].

In this Thesis, we applied the above-mentioned method to two separate datasets and constructed comparable call and text egocentric networks as well as combined egocentric networks from both calling and texting data. Next, from these egocentric networks, we constructed so-called *social signatures*. Social signatures quantify how people divide their communication effort among their alters. Then we studied the variation of social signatures across the population, over time, and across different channels of communication.

We observed that individuals have distinct social signatures that persist in time. This is true for all three types of social signatures–namely call signatures, text signatures, and mixed signatures. Similar observations were reported about social signatures made from the number of calls in Refs. [12, 14] and about email-based social signatures in Ref. [15]. In this Thesis, we confirmed this result for a large and demographically diverse sample of more than half a million individuals. The persistence of social signatures despite network turnover can indicate that the shape of social signatures is determined by stable individual characteristics. Ref. [14] reported that some behavioral traits can impact the stability of social signatures. Future research is needed on the roots of the persistence of the social signatures and also to reveal the consequences and causes of variation of social signatures across the population. For example, one can study the impact of demographic traits like gender and age on the shape of social signatures, since several studies have reported the impact of these traits on communication patterns [91, 92]. Moreover, since social relationships are crucial to the wellbeing of people, specific types of social signatures might associate with states such as loneliness and depression. Moreover,

unusual changes in the shape of social signatures over time can indicate changes in the behavioral traits of individuals and in cases might indicate social isolation. Future research can examine if there is any correlation between the shape or stability of social signature and mental wellbeing.

Another observation we reported in this Thesis is the similarity of individuals' social signatures across different communication channels of calls and texts. This similarity is unexpected considering the low values of Jaccard indices between the sets of alters in call and text egocentric networks. In other words, people use mobile-phone calls and text messages to contact different sets of people. However, they still distribute their calls among their call alters in the same way that they distribute their communication effort among their text alters. Why are egos' social signatures similar in shape across different channels? A possible hypothesis is the existence of an underlying complete social signature that captures the strength of all the relationships an ego maintains. Then the social signature through each communication channel shows an incomplete picture of that underlying signature but might still reflect some of its properties which makes the single-channel signatures similar. In order to find an explanation for the similarity of call and text signatures, we investigated the composition of mixed social signatures. However, we did not find any regularities or trend in the choice of channel concerning ranks of the alters in the mixed signatures. To understand this phenomenon, further research is needed on how people choose the channel of communication to maintain their social relationships.

The social signatures are constructed from the egocentric networks aggregated in a time window. However, none of the studies on the social signatures (including this Thesis) have investigated the impacts of the aggregation window on the properties of social signatures. A study similar to Ref. [82], which investigates the effects of aggregation window size on the structure of aggregated non-weighted communication networks, is needed to reveal the effect of the time window size on the properties of the social signatures.

References

[1] James S House, Karl R Landis, and Debra Umberson. Social relationships and health. *Science*, 241(4865):540–545, 1988.

[2] Julianne Holt-Lunstad, Timothy B Smith, and J Bradley Layton. Social relationships and mortality risk: a meta-analytic review. *PLoS medicine*, 7(7):e1000316, 2010.

[3] Roman M Wittig, Catherine Crockford, Julia Lehmann, Patricia L Whitten, Robert M Seyfarth, and Dorothy L Cheney. Focused grooming networks and stress alleviation in wild female baboons. *Hormones and Behavior*, 54(1):170–177, 2008.

[4] Mark S Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

[5] Sam GB Roberts and Robin IM Dunbar. The costs of family and friends: an 18-month longitudinal study of relationship maintenance and decay. *Evolution and Human Behavior*, 32(3):186–197, 2011.

[6] Debra L Oswald and Eddie M Clark. Best friends forever?: High school best friendships and the transition to college. *Personal Relationships*, 10(2):187–196, 2003.

[7] James Stiller and Robin IM Dunbar. Perspective-taking and memory capacity predict social network size. *Social Networks*, 29(1):93–104, 2007.

[8] George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81, 1956.

[9] Joanne Powell, Penelope A Lewis, Neil Roberts, Marta García-Fiñana, and Robin IM Dunbar. Orbital prefrontal cortex volume predicts social network size: an imaging study of individual differences in humans. *Proceedings of the Royal Society of London B: Biological Sciences*, 279(1736):2157–2162, 2012.

[10] Sam GB Roberts, Robin IM Dunbar, Thomas V Pollet, and Toon Kuppens. Exploring variation in active network size: Constraints and ego characteristics. *Social Networks*, 31(2):138–146, 2009.

[11] J-P Onnela, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and A-L Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, 2007.

[12] Jari Saramäki, Elizabeth A Leicht, Eduardo López, Sam GB Roberts, Felix Reed-Tsochas, and Robin IM Dunbar. Persistence of social signatures in human communication. *Proceedings of the National Academy of Sciences*, 111(3):942–947, 2014.

[13] Giovanna Miritello, Esteban Moro, Rubén Lara, Rocío Martínez-López, John Belchamber, Sam GB Roberts, and Robin IM Dunbar. Time as a limited resource: Communication strategy in mobile phone networks. *Social Networks*, 35(1):89–95, 2013.

[14] Simone Centellegher, Eduardo López, Jari Saramäki, and Bruno Lepri. Personality traits and ego-network dynamics. *PLOS ONE*, 12:1–17, 03 2017.

[15] Antonia Godoy-Lorite, Roger Guimerà, and Marta Sales-Pardo. Long-term evolution of email networks: statistical regularities, predictability and stability of social behaviors. *PlOS ONE*, 11(1):e0146113, 2016.

[16] Jean-Pierre Bocquet-Appel. When the world's population took off: the springboard of the neolithic demographic transition. *Science*, 333(6042):560–561, 2011.

[17] Charles H Feinstein. Pessimism perpetuated: real wages and the standard of living in britain during and after the industrial revolution. *The Journal of Economic History*, 58(3):625–658, 1998.

[18] Vaclav Smil. *Creating the twentieth century: Technical innovations of 1867-1914 and their lasting impact*. Oxford University Press, 2005.

[19] Petter Holme and Fredrik Liljeros. Mechanistic models in computational social science. *Frontiers in Physics*, 3:78, 2015.

[20] Thomas Haigh, Mark Priestley, and Crispin Rope. Los alamos bets on eniac: Nuclear monte carlo simulations, 1947-1948. *IEEE Annals of the History of Computing*, 36(3):42–63, 2014.

[21] Gregory Farrington. Eniac: The birth of the information age. *Popular Science*, 248(3):74–76, 1996.

[22] William F Brinkman, Douglas E Haggan, and William W Troutman. A history of the invention of the transistor and where it will lead us. *IEEE Journal of Solid-State Circuits*, 32(12):1858–1865, 1997.

[23] Robert R Schaller. Moore's law: past, present and future. *IEEE spectrum*, 34(6):52–59, 1997.

[24] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Computational social science. *Science*, 323(5915):721–723, 2009.

[25] Russell A Hill and Robin IM Dunbar. Social network size in humans. *Human Nature*, 14(1):53–72, 2003.

[26] Marc Keuschnigg, Niclas Lovsjö, and Peter Hedström. Analytical sociology and computational social science. *Journal of Computational Social Science*, 1(1):3–14, 2018.

[27] Jari Saramäki and Esteban Moro. From seconds to months: an overview of multi-scale dynamics of mobile telephone calls. *The European Physical Journal B*, 88(6):164, 2015.

[28] Anol Bhattacherjee. *Social science research: Principles, methods, and practices*. Global Text Project, 2012.

[29] H Russell Bernard, Peter D Killworth, and Lee Sailer. Informant accuracy in social-network data v. an experimental attempt to predict actual communication from recall data. *Social Science Research*, 11(1):30–66, 1982.

[30] Lauri Kovanen, Márton Karsai, Kimmo Kaski, János Kertész, and Jari Saramäki. Temporal motifs in time-dependent networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(11):P11005, 2011.

[31] Talayeh Aledavood, Eduardo López, Sam GB Roberts, Felix Reed-Tsochas, Esteban Moro, Robin IM Dunbar, and Jari Saramäki. Daily rhythms in mobile telephone communication. *PlOS ONE*, 10(9):e0138098, 2015.

[32] Lionel Tabourier, Alina Stoica, and Fernando Peruani. How to detect causality effects on large dynamical communication networks: a case study. In *Communication Systems and Networks (COMSNETS), 2012 Fourth International Conference On*, pages 1–7. IEEE, 2012.

[33] Ray M Chang, Robert J Kauffman, and YoungOk Kwon. Understanding the paradigm shift to computational social science in the presence of big data. *Decision Support Systems*, 63:67–80, 2014.

[34] Vincent D Blondel, Adeline Decuyper, and Gautier Krings. A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4(1):10, 2015.

[35] Jukka-Pekka Onnela, Jari Saramäki, Jörkki Hyvönen, Gábor Szabó, M Argollo De Menezes, Kimmo Kaski, Albert-László Barabási, and János Kertész. Analysis of a large-scale weighted network of one-to-one human communication. *New Journal of Physics*, 9.

[36] Márton Karsai, Mikko Kivelä, Raj Kumar Pan, Kimmo Kaski, János Kertész, A-L Barabási, and Jari Saramäki. Small but slow world: How network topology and burstiness slow down spreading. *Physical Review E*, 83(2):025102, 2011.

[37] James P Bagrow, Dashun Wang, and Albert-Laszlo Barabasi. Collective response of human populations to large-scale emergencies. *PlOS ONE*, 6(3):e17680, 2011.

[38] Mikko Kivelä, Raj Kumar Pan, Kimmo Kaski, János Kertész, Jari Saramäki, and Márton Karsai. Multiscale analysis of spreading in a large communication network. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(03):P03005, 2012.

[39] Giovanna Miritello, Esteban Moro, and Rubén Lara. Dynamical strength of social ties in information spreading. *Physical Review E*, 83(4):045102, 2011.

[40] Julián Candia, Marta C González, Pu Wang, Timothy Schoenharl, Greg Madey, and Albert-László Barabási. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015, 2008.

[41] Ye Wu, Changsong Zhou, Jinghua Xiao, Jürgen Kurths, and Hans Joachim Schellnhuber. Evidence for a bimodal distribution in human communication. *Proceedings of the National Academy of Sciences*, 107(44):18803–18808, 2010.

[42] Márton Karsai, Kimmo Kaski, Albert-László Barabási, and János Kertész. Universal features of correlated bursty behaviour. *Scientific Reports*, 2:397, 2012.

[43] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779, 2008.

[44] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[45] Francesco Calabrese, Giusy Di Lorenzo, and Carlo Ratti. Human mobility prediction based on individual and collective geographical preferences. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, pages 312–317. IEEE, 2010.

[46] Rainer Kujala, Talayeh Aledavood, and Jari Saramäki. Estimation and monitoring of city-to-city travel times using call detail records. *EPJ Data Science*, 5(1):6, 2016.

[47] Roberto Trasarti, Ana-Maria Olteanu-Raimond, Mirco Nanni, Thomas Couronné, Barbara Furletti, Fosca Giannotti, Zbigniew Smoreda, and Cezary Ziemlicki. Discovering urban and country dynamics from mobile phone data with spatial correlation patterns. *Telecommunications Policy*, 39(3-4):347–362, 2015.

[48] Thomas Louail, Maxime Lenormand, Oliva G Cantu Ros, Miguel Picornell, Ricardo Herranz, Enrique Frias-Martinez, José J Ramasco, and Marc Barthelemy. From mobile phone data to the spatial structure of cities. *Scientific Reports*, 4:5276, 2014.

[49] Balázs Cs Csáji, Arnaud Browet, Vincent A Traag, Jean-Charles Delvenne, Etienne Huens, Paul Van Dooren, Zbigniew Smoreda, and Vincent D Blondel. Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and Its Applications*, 392(6):1459–1473, 2013.

[50] Nathan Eagle and Alex Sandy Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006.

[51] Mika Raento, Antti Oulasvirta, Renaud Petit, and Hannu Toivonen. Contextphone: A prototyping platform for context-aware mobile applications. *IEEE Pervasive Computing*, 4(2):51–59, 2005.

[52] Arkadiusz Stopczynski, Vedran Sekara, Piotr Sapiezynski, Andrea Cuttone, Mette My Madsen, Jakob Eg Larsen, and Sune Lehmann. Measuring large-scale social networks with high resolution. *PlOS ONE*, 9(4):e95978, 2014.

[53] Leonhard Euler. Solutio problematis ad geometriam situs pertinentis. *Commentarii Academiae Scientiarum Petropolitanae*, 8:128–140, 1741.

[54] Jacob Levy Moreno, Helen Hall Jennings, et al. *Who shall survive?* Nervous and Mental Disease Publishing co., 1934.

[55] Mark Newman. *Networks: an introduction*. Oxford University Press, 2010.

[56] Fritz J Roethlisberger and William J Dickson. *Management and the Worker*. Harvard University Press, 1939.

[57] Guido Caldarelli. *Scale-free networks: complex webs in nature and technology*. Oxford University Press, 2007.

[58] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440, 1998.

[59] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[60] Paul Erdös and Alfréd Rényi. On random graphs, i. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959.

[61] Ithiel de Sola Pool and Manfred Kochen. Contacts and influence. *Social Networks*, 1(1):5–51, 1978.

[62] Jeffrey Travers and Stanley Milgram. The small world problem. *Psychology Today*, 1(1):61–67, 1967.

[63] Albert-László Barabási. *Network science*. Cambridge university press, 2016.

[64] John Guare. *Six degrees of separation: A play*. Vintage, 1990.

[65] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Internet: Diameter of the world-wide web. *Nature*, 401(6749):130, 1999.

[66] Charles Korte and Stanley Milgram. Acquaintance linking between white and negro populations: Application of the small world problem. *Journal of Personality and Social Psychology*, 15(2):101–108, 1970.

[67] Peter V Marsden and Karen E Campbell. Measuring tie strength. *Social Forces*, 63(2):482–501, 1984.

[68] FA Haight. Handbook of the poisson distribution. *Publications in operations research*, (11), 1967.

[69] Joao Gama Oliveira and Albert-László Barabási. Human dynamics: Darwin and einstein correspondence patterns. *Nature*, 437(7063):1251–1251, 2005.

[70] Albert-Laszlo Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.

[71] Zoltan Dezsö, Eivind Almaas, András Lukács, Balázs Rácz, István Szakadát, and A-L Barabási. Dynamics of information access on the web. *Physical Review E*, 73(6):066132, 2006.

[72] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. On the bursty evolution of blogspace. *World Wide Web*, 8(2):159–178, 2005.

[73] Alexei Vazquez, Balazs Racz, Andras Lukacs, and Albert-Laszlo Barabasi. Impact of non-poissonian activity patterns on spreading processes. *Physical Review Letters*, 98(15):158702, 2007.

[74] Dirk Brockmann, Lars Hufnagel, and Theo Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.

[75] R Dean Malmgren, Daniel B Stouffer, Andriana SLO Campanharo, and Luis A Nunes Amaral. On universality in human correspondence activity. *Science*, 325(5948):1696–1700, 2009.

[76] Mark Newman, Albert-Laszlo Barabasi, and Duncan J Watts. *The structure and dynamics of networks*. Princeton University Press, 2011.

[77] Petter Holme. Network reachability of real-world contact sequences. *Physical Review E*, 71(4):046119, 2005.

[78] Romualdo Pastor-Satorras and Alessandro Vespignani. *Evolution and structure of the Internet: A statistical physics approach*. Cambridge University Press, 2007.

[79] José Luis Iribarren and Esteban Moro. Impact of human activity patterns on the dynamics of information diffusion. *Physical Review Letters*, 103(3):038702, 2009.

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

[81] Jason Wiese, Jun-Ki Min, Jason I Hong, and John Zimmerman. You never call, you never write: Call and sms logs do not always indicate tie strength. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 765–774. ACM, 2015.

[82] Gautier Krings, Márton Karsai, Sebastian Bernhardsson, Vincent D Blondel, and Jari Saramäki. Effects of time window size and placement on the structure of an aggregated communication network. *EPJ Data Science*, 1(1):4, 2012.

[83] Marshall T Poe. *A History of Communications: Media and Society from the Evolution of Speech to the Internet*. Cambridge University Press, 2010.

[84] Robin IM Dunbar. Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22(6):469–493, 1992.

[85] Robin IM Dunbar. Neocortex size and group size in primates: a test of the hypothesis. *Journal of Human Evolution*, 28(3):287–296, 1995.

[86] Petter Holme and Jari Saramäki. Temporal networks. *Physics Reports*, 519(3):97–125, 2012.

[87] Petter Holme. Modern temporal network theory: a colloquium. *The European Physical Journal B*, 88(9):234, 2015.

[88] Sara Heydari, Sam GB Roberts, Robin IM Dunbar, and Jari Saramäki. Multi-channel social signatures and persistent features of ego networks. Proceedings in the Journal of Applied Network Science.

[89] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[90] Daniel Olguın Olguın and Alex Sandy Pentland. Sociometric badges: State of the art and future applications. In *Doctoral colloquium presented at IEEE 11th International Symposium on Wearable Computers, Boston, MA*, 2007.

[91] Kunal Bhattacharya, Asim Ghosh, Daniel Monsivais, Robin IM Dunbar, and Kimmo Kaski. Sex differences in social focus across the life cycle in humans. *Royal Society Open Science*, 3(4):160097, 2016.

[92] Sam BG Roberts and Robin IM Dunbar. Managing relationship decay. *Human Nature*, 26(4):426–450, 2015.