

Dynamics of High-Resolution Networks

Sekara, Vedran; Jørgensen, Sune Lehmann

Publication date:
2015

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

[Link back to DTU Orbit](#)

Citation (APA):

Sekara, V., & Jørgensen, S. L. (2015). Dynamics of High-Resolution Networks. Kgs. Lyngby: Danmarks Tekniske Universitet (DTU). (DTU Compute PHD-2015, Vol. 367).

DTU Library

Technical Information Center of Denmark

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

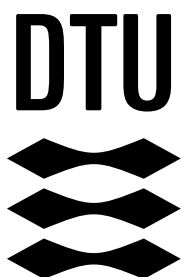
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Dynamics of High-Resolution Networks

Vedran Sekara

Advisor:
Sune Lehmann

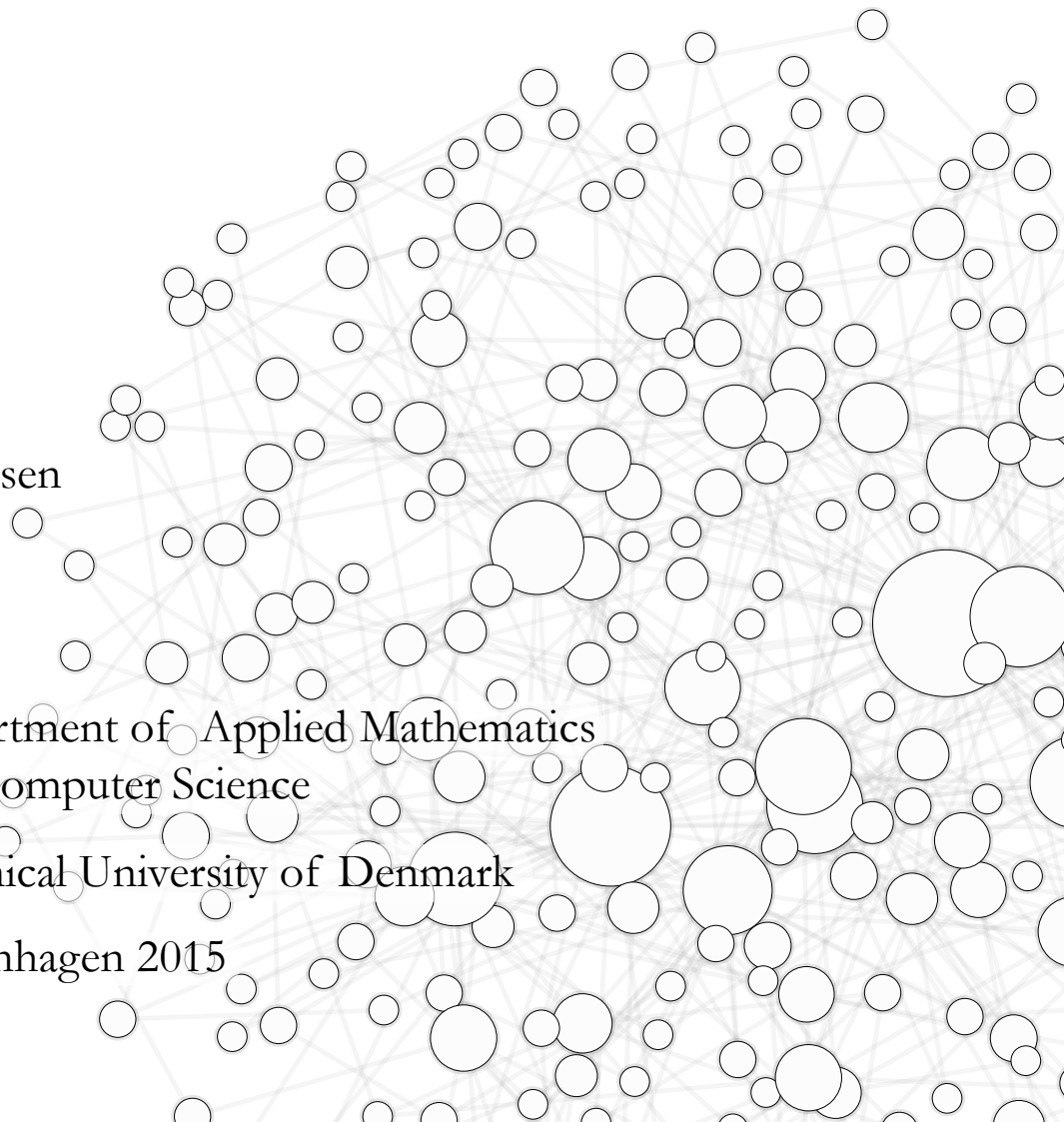
Co-Advisor:
Mogens Høgh Jensen



Department of Applied Mathematics
and Computer Science

Technical University of Denmark

Copenhagen 2015



Department of Applied Mathematics and Computer Science
Technical University of Denmark
2800 Kgs. Lyngby, Denmark

ISSN: 0909-3192

PHD-2015-367

Copyright © 2015 Vedran Sekara

Abstract

NETWORKS are everywhere. From the smallest confines of the cells within our bodies to the webs of social relations across the globe. Networks are not static, they constantly change, adapt, and evolve to suit new conditions. In order to understand the fundamental laws that govern networks we need new, highly detailed maps that uncover the interactions of all constituents, accurately and in a temporal manner. One of the hardest networks to understand, but also one of the most interesting ones, is the human social network. How do humans interact, form friendships, and spread information? And how are we all affected by an ever changing network structure? Answering these questions will enrich our understanding of ourselves, our organizations, and our societies. Yet, mapping the dynamics of social networks has traditionally been an arduous undertaking. Today, however, it is possible to use the unprecedented amounts of information collected by mobile phones to gain detailed insight into the dynamics of social systems. This dissertation presents an unparalleled data collection campaign, collecting highly detailed traces for approximately 1 000 people over the course of multiple years. The availability of such dynamic maps allows us to probe the underlying social network and understand how individuals interact and form lasting friendships. More importantly, these highly detailed dynamic maps provide us new perspectives at traditional problems and allow us to quantify and predict human life.

Resumé (summary in Danish)

Vi er omgivet af netværk. Fra de inderste dele af vores celler til de globale venskabsbånd mellem mennesker. Netværk er ikke statiske, men under konstant evolution for at tilpasse sig vekslende miljøforhold. Vi har derfor brug for kort—meget detaljerede og præcise kort—for at forstå de bagvedliggende mekanismer og grundlæggende love, der styrer netværk. Et af de mest komplekse, men samtidig mest spændende netværk er båndet mellem mennesker, det vi i daglig tale kalder sociale netværk. Hvordan interagerer mennesker? Hvordan danner vi venskaber? Og præcis hvordan deler vi viden? Svarene på disse spørgsmål vil give os en bedre forståelse af hvordan vores liv, organisationer, og samfund hænger sammen. En kortlægning af disse dynamikker er dog ikke et trivielt anliggende og har tidligere vist sig at være meget vanskelig. Grundet teknologiske fremskridt er det dog i dag muligt at indsamle uanede mængder af information gennem vores mobiltelefoner. Denne PhD-afhandling præsenterer et enestående eksperiment hvor vi ved brug af moderne *smartphones* har indsamlet detaljeret adfærd for 1 000 mennesker over flere år. Tilgængeligheden af sådanne dynamiske kort, giver os en enestående mulighed for at kortlægge og forstå hvordan individer interagerer og danner grupper. Disse kort giver os også mulighed for at kvantificere vores adfærdsmønstre og lader os i sidste ende sætte tal på hvor forudsigelige vores liv er.

Contents

Contents	v
List of Figures	vii
Papers	ix
Preface	xi
Acknowledgments	xiii
1 Introduction	1
1.1 Aim	3
1.2 Outline	4
2 Complex networks	5
2.1 Statistical measures	8
2.1.1 Degree distribution	9
2.1.2 Mixing and degree correlations	10
2.1.3 Shortest paths, diameter and betweenness	11
2.1.4 Navigability	12
2.1.5 Clustering	13
2.2 Network models	14
2.2.1 Random Networks	15
2.2.2 Small-world networks	16
2.2.3 Scale-free networks	17
2.3 Network resilience	18
2.4 Network growth	20
2.5 Community structure	22
2.6 Temporal networks	26
2.6.1 Examples of temporal networks	27
2.6.2 Properties of temporal networks	29

3	Measuring Networks	35
3.1	Proxy for human life	36
3.2	Data collection endeavors	37
3.3	Copenhagen Networks Study	38
	<i>Summary of Paper I</i>	39
3.4	Inferring relations	41
	<i>Summary of Paper II</i>	43
4	Understanding social systems	45
4.1	Uncovering routine	46
4.2	Social groups	48
4.2.1	Network representations	50
4.2.2	Gatherings	50
4.2.3	Dynamic communities	53
	<i>Summary of Paper III</i>	56
5	Summary	61
5.1	Concluding remarks	62
	References	63
	Publications	I
	<i>Paper I</i>	III
	<i>Paper II</i>	XXVII
	<i>Paper III</i>	XXXV
	Understanding scientific collaboration	LXXXIII
	<i>Paper IV</i>	LXXXV

List of Figures

2.1	A network	9
2.2	Network measures	11
2.3	Network models	15
2.4	Network degree distributions	19
2.5	Network resilience	20
2.6	Community structure in networks	23
2.7	Temporal network representations	29
2.8	Network growth	31
2.9	Time-respecting paths	31
3.1	Growth of mobile technology	37
3.2	Dynamics of face-to-face interactions	40
3.3	Perspectives of a network	41
3.4	Heuristics of removing ties	44
4.1	Regular patterns within a network	46
4.2	Physical proximity networks	51
4.3	Gathering dynamics	52
4.4	Gathering boundaries	53
4.5	Cores summarize social context	58

Papers

This dissertation is based on the following papers:

- I Stopczynski, A., **Sekara, V.**, Sapieżyński, P., Cuttone, A., Madsen, M. M., Larsen, J. E., and Lehmann, S. (2014). *Measuring Large-Scale Social Networks with High Resolution*. PLoS ONE, 9(4):e95978.
- II **Sekara, V.**, and Lehmann, S. (2014) *The Strength of Friendship Ties in Proximity Sensor Data*. PLoS ONE, 9(7):e100915.
- III **Sekara, V.**, Stopczynski, A., and Lehmann, S. (2015). *Uncovering Fundamental Structures of Dynamic Social Networks*. In submission.

Other papers by the author not included in the main storyline of the dissertation are:

- IV **Sekara, V.**, Sinatra, R., Deville, P., Ahnert, S., Barabási, A.-L., and Lehmann S. (2015) *The Chaperone Phenomenon in Science*. In preparation.

All publications are listed in the given order at the end of the dissertation.

*“The roots of education are bitter,
but the fruit is sweet.”*

— Aristotle

Preface

THIS is a story of network science, not the full unabridged story, as it is way too long to be condensed into just one dissertation. Rather it is my story. A journey that took me from physics into a new science, network science.

A physicist by training, I was raised in the fabled hallways and auditoriums of the Niels Bohr Institute. A place which in the 1920’s and 30’s was the center for quantum and atomic physics. Around the institute hung these old photographs depicting the countless great minds of physics who had previously visited Niels Bohr. Heisenberg, Dirac, Pauli, Born, Gamow, Landau, Ehrenfest, Delbrück, the list goes on, had all been sitting in the same auditoriums which we were now lectured in. Here, my fellow students and I were trained in the mystical art of intuition, stringent logics and were presented with a countless number of equations, too many to memorize.

After finishing my studies I was not really eager to get a *real* job, I felt that there was so much more to learn. So I decided to pursue a PhD where I could delve into the physics of complex systems, but I was unsure about which academic institution I should study at. I hastily sent out three applications to different PhD schools and to my great surprise I was accepted by all three. After great consideration I had finally settled on one of the programs, when suddenly my Master’s thesis advisor, Mogens Høgh Jensen, told me that I had yet another option. He had just talked to a smart and trustworthy young professor who had recently received a large grant to study social systems with smartphones—an ambitious but *slightly crazy* project, I was told. Nevertheless, I went to talk to the guy and he somehow convinced me to decline all other offers, and continue my studies under his guidance.

So here I am, a physicist that has devoted three years of his life trying to untangle the unbelievably complex web of relations between people and understand how our society works. To my great amusement I am not the first physicist to invade somebody else’s field, as Duncan Watts so eloquently put it:

“Physicists, it turns out, are almost perfectly suited to invading other people’s disciplines, being not only extremely clever but also generally

much less fussy than most about the problems they choose to study. Physicists tend to see themselves as lords of the academic jungle, loftily regarding their own methods as above the ken of anybody else and jealously guarding their own terrain. But their alter egos are closer to scavengers, happy to borrow ideas and techniques from anyone if they seem like they might be useful, and delighted to stomp all over someone else's problem. As irritating as this attitude can be to everybody else, the arrival of physicists into a previously non-physics area of research often presages a period of great discovery and excitement. Mathematicians do the same thing occasionally, but no one descends with such fury and in so great a number as a pack of hungry physicists, adrenalized by the scent of a new problem."

— Duncan J. Watts, *Six Degrees*

The first time I came across this quote I was attending an interdisciplinary conference and could not stop laughing. Watts is correct, as can be inferred from the list of references at the back of this dissertation, a majority of the listed authors are other physicists, who in their quest for new problems decided to cannibalize somebody else's field. If you look closer, however, you will see that the physicists are not there alone, listed alongside them are: mathematicians, computer scientists, biologists, ecologists, sociologists, economists, psychologists, anthropologists, linguists and many others. During my PhD studies I have had the pleasure and pain of working with many of them. As it turns out, a multidisciplinary collaboration can be very rewarding in terms of new ideas, but also highly frustrating as our scientific discourses are so very different. I recall with great delight the many occasions, where I wholeheartedly attempted at conveying complicated mathematical concepts to researchers from the social sciences, and other great moments where they tried to educate me in the ways of a social scientist. Nevertheless, my journey through the PhD has been an educational, challenging and rewarding experience.

Acknowledgments

THANK god its over! I am eternally indebted to a great deal of people, who over the years have graciously enlightened my life with dirty jokes, inspiration, encouragement, and scientific advice.

Sune Lehmann for giving me an unparalleled opportunity to play with hardware worth 3 million Danish kroner, and for the countless unrestrained discussions about life, science, and just about anything else. This journey was more fun than I could have imagined.

Albert-László Barabási for letting me visit his lab and Roberta, Pierre, Marc, Gabriele, and all the others at the CCNR who made my stay in the US memorable. This stay would not have been possible without generous funding from Knud Højgaards Foundation, the Niels Bohr Foundation, the Oticon Foundation, and the Otto Mønsted Foundation. Moreover, my PhD has been supported by a Young Investigator Grant from the Villum Foundation awarded to Sune Lehmann.

A great thanks goes to my intrepid labmates, especially Arkadiusz Stopczynski and Piotr Sapieżyński who made my long days at the office quite enjoyable.

My dear friends and my housemates from Poco Loco. It pains me not to list you all, but I tried and the list got unbelievably long. Thank you all for keeping me sane and supplying me with beer, food, chocolate, and good advice when I needed it the most.

I would to thank my family: mother, father, and sister; thank you for your unrelenting optimism and support. A special thanks goes to the Skadegaard Thorsen family for providing me with a second home.

Most importantly, thank you Tess. Thank your for showering me with unlimited amounts of love. You are intelligent, beautiful, independent, charming, and an exceptionally skilled proofreader. I love you more than words can express. I will always strive to maintain your love and affection.

And thank you dear reader for picking up this dissertation and at least making it this far. Thank You.

1

Introduction

THE most striking aspect of physics is the simplicity of its Laws. Maxwell's equations, Newton's laws of motion, Hamiltonian mechanics, and Schrödinger's equation. Everything is simple, neat, and beautiful. However, venturing outside the safe confines of a physics classroom and out into the world, everything is not as promised—the world is astonishingly complex. Everywhere we look we see truly magnificent examples of complex phenomena: from the delicate ridges of Saharan sand dunes formed by the wind, to the intricate ecological organization of species, and the swarm-like behavior of bird flocks (Goldenfeld and Kadanoff, 1999). The traditional physicists approach in studying such complex systems is to first reduce them to their constituent components and understand how they individually work. From here we can then piece together the more complex phenomena. Applying this reductionist reasoning, physics has been quite successful over the years, in describing the fundamental behaviors and interactions of particles, up to the scale of atoms. Put a few atoms together, however, and everything becomes impossibly complex again, that is why chemistry is a science in itself and not just a branch of physics, biology cannot just be reduced to chemistry and medical sciences are more than just applied biology (Watts, 2004). There is a hierarchy of complexity, at each level entirely new properties appear and the understanding of these behaviors requires new endeavors of research (Anderson et al., 1972).

Complex systems are hard to solve because they are made up of many constituent parts that do not add up in any straightforward manner, meaning that

their collective behavior is not a simple combination of individual behavior. On the contrary, the components interact in completely new and often unforeseen ways. Take our genome as an example, even though we are 99.9% similar to every other human we still exhibit diverse phenotypes such as eye color, hair color, skin color, height, body type, etc. In reality we humans consist of roughly 20 000 genes, similar in number to many insects, worms, and fish, but what makes us different from them? Genetic variation clearly does not stem from the raw number of genes, rather it arises from their interconnectedness, as genetic traits are not expressed by single genes but in a collective fashion. In a similar manner yours, and my brain consist of roughly 100 billion interconnected neurons similar in numbers to *William Shakespeare's* and *Leonardo da Vinci's*, but when did we last write our *Hamlet* or paint our *Mona Lisa*? This non-trivial phenomena where two otherwise identical entities can produce different types of behavior is what sets complex systems apart from simpler systems.

With this in mind, what hope do we have of understanding systems such as our society? If each individual is composed of complex entities, placing humans together would surely produce even more stunning examples of complexity. As it turns out, when many humans come together it is possible to ignore the complicated details such as unique characteristics and personalities of individuals and extract basic organizing principles from our aggregated patterns. In fact, while we cannot learn much about society from any single individual, when many of us come together collective patterns emerge.

Behind each complex system there is a intricate wiring diagram, that defines the interactions between components—we call this a network. Networked systems differ greatly from one another. Nodes, for example, can represent entities from tiny molecules to Internet routers, or humans, while links can denote chemical reactions governed by quantum mechanics, physical cables laid down by people, or social relations between families, friends, and professionals. The processes that have shaped these systems are also of fundamentally different nature, while metabolic networks have been shaped by billions of years of evolution, the Internet has been collectively built during the last half century, and our social networks are deeply rooted in cultures and norms dating back thousands of years. Given this diversity in size, origin, history, and evolution one would expect that the underlying networks behind these systems would greatly differ. Yet, the architecture of all these diverse systems is quite similar and allows us to formulate a shared set of mathematical tools. This principle of universality is what makes networks special, as fundamentally different systems observed within the domains of biology, physics, mathematics, economy, sociology, computer science, etc. can all be described by one single mathematical formalism. Thus, if we desire a deeper understanding of our complex world, networks are an excellent place to start.

1.1 Aim

In order to understand complex systems and describe their behaviors we first need to obtain a map of the system. This map should specify the wiring diagram of the network. For a social system this requires knowing the list of your friends, their friends, their friends' friends and so on. In our genome this corresponds to knowing all interactions between genes, proteins, and their associated interrelations to external environmental factors. A feat we, in spite of countless modern medical breakthroughs, are not even close to achieving. Even worse, as none of these systems are static by nature, this calls for dynamic maps. Thus, in the example of social systems it is not sufficient to know who is friends with whom, but also when they interact. Yet another important aspect is to know how entities interact, particularly through which types of relations. For example do people meet up and talk face-to-face or do they communicate via e-mails, phone calls or text messages? This is an important piece of information as messages can propagate with different speeds depending on the type of tie.

The main goal of this dissertation is to take advantage of recent technological advances in order to collect a dynamic map of unprecedented quality. We focus on social systems, where this map will give us unparalleled opportunities to study the intricate web of human relations and allow researchers to gain novel insights about the social fabric itself. Even though we focus on mapping and understanding social systems, the universality of networks will enable us to generalize some of our findings to other domains of network science.

In order to construct a comprehensive picture of the social network we first need to map the web of interactions. As human dynamics unfold over many timescales, from transient encounters to long lasting relationships, it is important that we sample the network at appropriate levels of detail. Specifically, this thesis aims to record the network of social interactions with very high resolution, both in terms of temporal sampling and number of recorded communication channels. We record data using state-of-the-art smartphones as sensors¹ as they enable us to capture multiple communication channels using only a single device.

Collecting social network data with mobile devices is still a relatively new field, so first we need to calibrate our instrument of measurement before we can apply it to gain theoretical insight. The questions we aim to answer are: What does an electronic “click” entail, how can they effectively be used to understand human relationships, and what are the shortcomings of using smartphones as sensors?

After having laid the groundwork in collecting data and calibrating our instruments we aim to tackle one of the fundamental problems of behavioral

¹Also denoted as sociometers in other parts of the scientific literature see Pentland (2008).

science, are humans predictable and to what degree? We will look at human behavioral patterns and extract regularities which we can use to quantify the actions of each individual. Life is imbued with regularities in many of its facets, the patterns we aim to understand are geospatial and social. We will look into the two aspects of daily life and try to understand them separately and combined. In doing so we will formulate new solutions to the problem of identifying communities within networks.

1.2 Outline

This dissertation regards the dynamics of high-resolution networks. It is divided into five main parts. Chapter 1 serves as a general introduction into the field of complex systems and explains why networks are important. Chapter 2 is directed at the growing field of complex networks and gives a theoretical overview of important concepts and models. As there are already many good reviews on networks science, chapter 2 tells a story that binds the dissertation together rather than giving a comprehensive review. For general audiences I would specifically recommend the books by Barabasi (2002), Buchanan (2003), and Watts (2004). Curious scientific readers should read the papers by Newman (2003), Newman (2005), Boccaletti et al. (2006) and Fortunato (2010), and the books by Newman et al. (2006) and Easley and Kleinberg (2010). The central parts of this dissertation are located in chapters 3 and 4 which chronicle my research as a PhD student. Chapter 3 describes a unparalleled data collection campaign and investigates how we can use the collected traces to infer real social interactions. Chapter 4 digs into the regularities of human behavior and presents a novel approach to understanding and extracting dynamic communities. The last chapter (chapter 5) briefly summarizes the main findings of the dissertation. All papers are listed at the back along with an additional publication that is not included in the main storyline of the dissertation.

Without further ado we will continue with a quote:

"Sal, we gotta go and never stop going 'till we get there."

"Where we going, man?"

"I don't know but we gotta go."

— Jack Kerouac, *On the Road*

2

Complex networks

NETWORKS are everywhere. In fact, we spend our entire lives being entangled in and navigating numerous complex networks. Our social ties form different webs: networks of friends, family, coworkers, or sexual partners. We keep in touch by communicating through a variety of technological networks such as: telephones, the Internet, and the World Wide Web². Infrastructural networks facilitate the transportation of electricity via power-lines, while an underground network of pipes supply us with water. We travel on transportation networks, be it roads, trains, ships, or airplanes. Our biology is governed by networks ranging from microscopic regulatory networks inside our cells to macroscopic ecological webs between species.

Networks have always surrounded us, and always will. In modern history they were first noticed by Leonhard Euler in 1735, when he at the age of 28 solved the *Königsberg bridge problem* by reducing it to a set of nodes and links, thus laying the foundations for graph theory³. During the twentieth, and the beginning of the twenty-first century the field of network theory has expanded and developed into a substantial body of knowledge. Emphasizing the strength of Euler's approach and illustrating that, as mathematical objects, networks are not restricted to only represent bridges between landmasses, but can be

²The distinction between the two latter is that the Internet denotes the physical structure of routers and servers, while the World Wide Web covers connections of web pages.

³The difference between *graphs* and *networks* is subtle, but where a graph is an abstract mathematical object a network often refers to an actual physical system.

applied in order to understand the features of a variety of complex systems. As such, researchers have applied network theory in order to describe a plethora of real world systems which can be loosely classified into four categories: social, information, technological, and biological systems.

Social networks A social network consists of a set of individuals with some interaction pattern between them (Wasserman, 1994; Scott and Carrington, 2011). Consequently, this representation can be applied to describe friendships between individuals (Rapoport and Horvath, 1961), communication between terrorists (Krebs, 2002), patterns of sexual contacts (Liljeros et al., 2001; Bearman et al., 2004), business relations between companies (Mariolis, 1975; Davis et al., 2003), and intermarriages between families (Padgett and Ansell, 1993). The field of social network analysis was pioneered by Jacob Moreno who in the 1930's used *sociograms* to study interpersonal relationships within small groups (Moreno and Jennings, 1934). A few years later Davis et al. (1941) studied the social circles of women in an unnamed city in the southern part of USA, and Rothlisberger and Dickson (1939) studied relations between factory workers in Chicago. Traditional studies of social networks have been limited to very small sample sizes because the object of analysis—social ties—are notoriously hard to observe. With the exception of a few ingenious studies such as Milgram's *small-world* experiment (Milgram, 1967), researchers have mostly relied on questionnaires, interviews, and self reports (Watts, 2007). The problem with survey data, however, is that it suffer from small sample sizes, cognitive biases, errors of perception, and ambiguities such as how respondents might define a friendship (Wuchty, 2009). In addition, surveys are very labor intensive and mostly comprised of one-time snapshots, thereby being unable to capture behavioral patterns over extended periods of time. Although much effort has been put into resolving these issues it is generally assumed that surveys contain uncontrollable errors (Marsden, 1990).

With the advent of the Internet, new relatively reliable data sources have appeared. At the end of the twentieth century researchers had easy access to collaboration networks between scientists (Newman, 2001), work relations between movie actors (Adamic and Huberman, 2000; Amaral et al., 2000), co-appearances of comic book heroes (Alberich et al., 2002), and telephone calls between individuals (Aiello et al., 2000). These were followed by datasets regarding: e-mail communication (Ebel et al., 2002; Newman et al., 2002; Eckmann et al., 2004; Kossinets and Watts, 2006), instant-messaging networks (Leskovec and Horvitz, 2008), and online social networks (Lewis et al., 2008; Aral and Walker, 2012). As the technological evolution continued, additional data sources became available, such as mobile phone calls between individuals (Onnela et al., 2007; Miritello et al., 2013), and time resolved contact patterns

between teams (de Montjoye et al., 2014), students (Eagle and Pentland, 2006; Stehlé et al., 2011b; Stopczynski et al., 2014a), and patients and health-care workers (Isella et al., 2011).

Information networks Information networks represent the exchange of *knowledge* between parties. A classical example is the web of citations between academic papers, where articles are nodes and citations embody the transfer of information. This is a special kind of network, because new papers can only cite already written papers all links will point backwards in time and will rarely, if ever, be reciprocated. The great advantage of citation networks is the accuracy and abundance of data stretching over multiple decades. Lotka (1926) applied this data to investigate the productivity of scientists and de Solla Price (1965) pointed out large differences between papers.

A different information network is the World Wide Web, currently consisting of at least 47.45 billion pages⁴ connected via hyperlinks. Its structure has been extensively studied by Albert et al. (1999), Kleinberg et al. (1999), Barabási et al. (2000), and Adamic and Huberman (2000) to name a few.

Other notable examples of information networks are citations between U.S. patents (Jaffe and Trajtenberg, 2002), spreading of mobile phone viruses (Wang et al., 2009), resubmission networks between academic journals (Calcagno et al., 2012), networks of cultural history (Schich et al., 2014), and semantic relations between words (Sigman and Cecchi, 2002).

Technological networks These networks are typically man-made and constructed for the distribution of a specific commodity. A well known technological network is the Internet, organized by the physical connection of routers. Using special programs to track the paths of data packets Faloutsos et al. (1999) mapped the Internet and extensively studied its properties. Other examples of technological networks are airline traffic networks (Amaral et al., 2000), networks of roads (Rosvall et al., 2005; Porta et al., 2006), railways (Latora and Marchiori, 2002; Sen et al., 2003), and electric power grids (Watts and Strogatz, 1998). Although they are naturally occurring, rivers can also be considered as a form of distribution network (Maritan et al., 1996; Dodds and Rothman, 1999).

An interesting feature of technological networks is that their properties to a large degree are governed by space and geography (Yook et al., 2002; Daqing et al., 2011). Links between subunits, be they electrical transformer stations or cities, are influenced by economic concerns, technological challenges, and geographical factors. As such the interplay of external factors that affect the formation of links in spatial embedded networks is not yet fully understood.

⁴As of Sunday 29th March 2015 (<http://www.worldwidewebsize.com>).

Biological networks Each living organism is embedded in a multitude of biological webs. Large ecosystems are formed from the complex interactions of species (Baird and Ulanowicz, 1989; Cohen et al., 1990), and their topologies have been comprehensively studied by Dunne et al. (2002), Williams et al. (2002), and Jordano et al. (2003) among many others. Looking at smaller length scales our brains are composed of billions of interconnected neurons forming neural networks (Sporns, 2002). Due to the staggering number of nodes we have yet to see a complete picture of the human brain, but White et al. (1986) have succeeded in mapping the entire neural network of the nematode (roundworm) *C. elegans*.

Even the regulatory system that allows organisms to sense and respond to changing environmental conditions form a web. Here the expression of a gene, i.e. the process of transcribing and translating genetic material into a protein, is controlled by the production of other proteins encoded by different genes; as such our entire genome forms a network⁵. Representing genes as nodes and their interactions with protein as edges Shen-Orr et al. (2002) and Guelzim et al. (2002) among others, have studied the statistical properties of regulatory networks. Elowitz and Leibler (2000) have even been successful in creating a simple synthetic regulatory motif consisting of three genes, which they have implanted in living *E. coli*.

Within cells, proteins mechanically cooperate with other proteins forming an additional network of protein-protein interactions. A network that has been extensively studied by many authors (Uetz et al., 2000; Ito et al., 2001; Jeong et al., 2001; Maslov and Sneppen, 2002; de Lichtenberg et al., 2005). At even smaller scales we can represent metabolic reactions of substrates and products as networks, linking two metabolites together if they are part of the same reaction. The properties of such metabolic networks have been studied by many scientists, including Jeong et al. (2000), Wagner and Fell (2001), Ravasz et al. (2002) and Stelling et al. (2002).

The above mentioned networks vary in size, function, and nature representing connected systems on the order of a few nodes to billions of connected entities. While previously it was difficult to acquire such data, the Internet revolution has made it relatively easy. Nowadays, we can collect networked datasets from *Twitter*, the blogosphere, or from any of the countless available network databases with little effort. However, in order to understand and describe these rich patterns of interaction we need mathematical tools that embrace the apparent diversity of complex systems. Figure 2.1 illustrates the problem we are facing. It shows a relatively small network consisting of 263 individuals, linked together if

⁵In certain cases, if the concentration of a specific protein is high, the protein can even regulate its own gene thus forming self loops in the network.

they share a friendship. From visual inspection we can see that some individuals have more friends than others, that nodes aggregate and form groups, and that the structure of the social network is inhomogeneous. While the human eye is good at revealing structure, even this small network poses a problem as the tangle of links makes it hard to describe the network in a satisfactory manner. For larger networks visual inspection is completely out of the question and we need statistical tools in order study their properties.

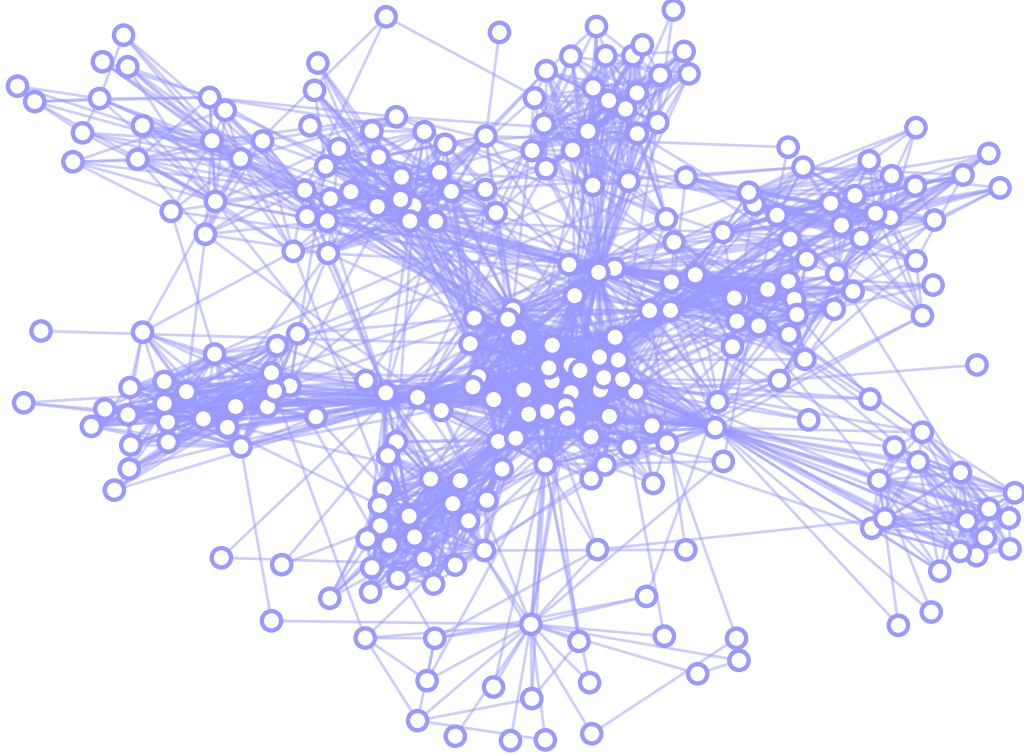


Figure 2.1: A network showing the friendship structure between 263 students at a Danish University. Each links symbolizes an online friendship. Data is collected from *Facebook* and supplied by Snorre Ralund, University of Copenhagen.

2.1 Statistical measures

Each network is composed of a number of nodes n wired together via m links. Their connections can be represented in a variety of ways, with the most simple being binary, describing whether the link is present or not. Such wiring schemes are defined by the adjacency matrix

$$A_{ij} = \begin{cases} 1 & \text{if nodes } i \text{ and } j \text{ share a connection,} \\ 0 & \text{otherwise.} \end{cases} \quad (2.1)$$

Other ways to represent edges is with associated weights that record their strength relative to each other, e.g. describing the number of calls between persons i and j , or by allowing multiple types of edges to connect nodes, letting i and j interact across multiple network layers. In addition, each tie can have an intrinsic direction, indicating whether i contacted j or vice versa. In order to keep things as simple as possible, we will focus exclusively on single layer networks comprised of undirected and unweighted links, however, a majority of the statistical measures can be generalized for the case of directed, weighted, and multiplex networks (Barrat et al., 2004; Newman, 2004a; Boccaletti et al., 2006; De Domenico et al., 2013b).

2.1.1 Degree distribution

The degree k_i denotes the number of links or edges that are connected to node i , as a result k_i can also be regarded as the number of neighbors of i (Figure 2.2 a). It can easily be calculated by taking the corresponding row (or column) sum of the adjacency matrix⁶. Because nodes in a network typically do not have the same number of links, e.g. humans have different numbers of friends, we characterize a system using the degree distribution $P(k)$. The probability function $P(k)$ thus denotes the probability that a randomly chosen node has exactly k links.

To gain a qualitative understanding of the distribution we can look at its moments. The q^{th} moment of the degree distribution is defined as

$$\langle k^q \rangle = \sum_k k^q P(k) = \int_k k^q P(k) dk. \quad (2.2)$$

In analytical calculations it is often more convenient to assume that degrees can take any positive real values; thus the second definition. The lower moments contain important physical interpretations about the distribution.

$q = 0$: The zeroth moment, $\langle k^0 \rangle$, describes the mass of the distribution, where $\sum_k P(k) = 1$.

$q = 1$: The first moment, $\langle k \rangle$, is the average of the distribution, i.e. the average degree.

$q = 2$: The second moment, $\langle k^2 \rangle$, describes the variance of the distribution $\sigma^2 = \langle k^2 \rangle - \langle k \rangle^2$, measuring the spread in the number of connections.

⁶If the network is undirected the row and column sums will be identical ($k_i = \sum_j A_{ij} = \sum_i A_{ij}$), otherwise the sums denote the number of outgoing ($k_i^{\text{out}} = \sum_j A_{ij}$) and the number of incoming connections ($k_i^{\text{in}} = \sum_{i'} A_{i'i}$).

$q = 3$: The third moment, $\langle k^3 \rangle$, measures the skewness of the distribution, revealing how symmetric $P(k)$ is around the average. A symmetric distribution has zero skewness.

As we will see later in Section 2.2, the degree distribution plays a central role in characterizing the structure of a network.

2.1.2 Mixing and degree correlations

Exploring the topology of networks more rigorously, we can look at which pairs of nodes tend to link up to each other. Since many networks consist of different types of nodes, the probability of them linking up will depend on the type. A good example is a food web where nodes represent species and links denote their predator-pray relations (Cohen et al., 1990). Here, there are many edges linking herbivores and plants and herbivores and carnivores together, while direct relations between herbivores and herbivores and carnivores and plants are unlikely (Williams and Martinez, 2000).

Social networks display a different kind of connection pattern, called assortative mixing or homophily, where we tend to associate with people that are similar to ourselves. A classical example is the choice of partner, where ethnically distinct groups tend to mix together, preferentially picking partners with the same ethnic background (Kalmijn, 1998). Because similarity breeds connection homophily has also been observed for characteristics such as age, religion, education, occupation, and gender (McPherson et al., 2001).

It is particularly interesting to look at mixing patterns between the degree of nodes and ask whether high degree nodes link up with other high degree nodes, or if they prefer to attach to low degree ones. The chance that a node with degree k is connected to a node of degree k' is given by the conditional probability $P(k'|k)$. For nodes with degree k we can calculate the average degree of their nearest neighbors as

$$\langle k_{nn} \rangle(k) = \sum_{k'} k' P(k'|k). \quad (2.3)$$

If there is a correlation between the degree of nodes, such that high degree nodes link up to other high degree nodes, then we call it assortative mixing and $\langle k_{nn} \rangle(k)$ will be an increasing function of k . The reverse situation where $\langle k_{nn} \rangle(k)$ decreases as a function of k signals that high degree nodes prefer to connect to low degree nodes. Such mixing is denoted as being disassortative. As it turns out, we observe both situations in networks (Newman, 2002). A majority of social networks appear to be assortative, while information, technological and biological appear to be disassortative.

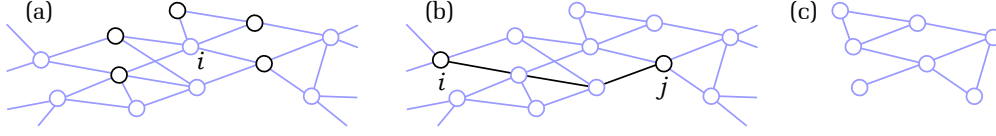


Figure 2.2: Network measures. (a) Node i has a degree of 4 and its neighborhood is marked in black. (b) The shortest path between nodes i and j has length 3. There are four such paths, one of them is marked in black. (c) Transitivity in the network is $T = 3 \cdot 2 / 17 \approx 0.353$ and average clustering is $C = 1/7 \cdot (1 + 1/3 + 1/3 + 1/3 + 1 + 1/6 + 0) \approx 0.452$.

2.1.3 Shortest paths, diameter and betweenness

Knowing the shortest paths plays a vital role in navigating and communicating within networks. Suppose you want to travel from Copenhagen to Boston, knowing the optimal path would save you considerable amounts of time. Distance between nodes is measured as the number of links one needs to traverse in order to get from node i to node j , this is also known as the geodesic distance. Note that there may be multiple shortest paths connecting two nodes, meaning that the shortest path between Copenhagen and Boston (because there are no direct flights) can take you through Frankfurt, London, Reykjavik, or some other city (Figure 2.2 b). Watts (1999) defined the typical separation between two nodes in a network as the average shortest path length, calculated as the mean geodesic distance (d_{ij}) between all pairs of nodes:

$$l = \frac{1}{n(n-1)} \sum_{i,j, i \neq j} d_{ij}. \quad (2.4)$$

Consequently the diameter of a network is defined as the longest geodesic path between any two nodes. There is, however, a problem with this definition as l diverges if there are disconnected nodes or components. Latora and Marchiori (2001) proposed an alternative and defined the average efficiency or *harmonic mean* as

$$l^{-1} = \frac{1}{n(n-1)} \sum_{i,j, i \neq j} d_{ij}^{-1}. \quad (2.5)$$

If d_{ij} diverges l^{-1} will still be well defined. The average distance tells us something about the size of a network, but it lacks any information about which paths are important. Communication between two non-neighboring nodes i and j , depends on other nodes on the paths that connect them. In this regard, we can measure the importance of a node by counting the number of shortest paths that go through it. As such the betweenness (b) of node u is defined as

$$b_u = \sum_{i,j, i \neq j} \frac{\sigma_{i,j}(u)}{\sigma_{i,j}}, \quad (2.6)$$

where $\sigma_{i,j}$ is the total number of shortest paths that connect nodes i and j , and $\sigma_{i,j}(u)$ is the number of these paths that go through node u . As an analogy to node betweenness we can similarly define the betweenness of an edge as the number of shortest paths that utilize a specific tie (Costa et al., 2007). Along with the degree, betweenness measures the centrality of a node (or an edge) within a network, consequently the concept can be applied in order to understand the vulnerability of networks, quantifying the impact of structural failures on network performance (Holme et al., 2002; Newman, 2003). It is possible to define centrality in different ways, as such variations of centrality include: closeness centrality (Sabidussi, 1966), straightness centrality (Vragović et al., 2005), Katz centrality (Katz, 1953), Eigenvector centrality (Newman, 2008), and information centrality (Porta et al., 2006).

2.1.4 Navigability

The above measures tell us which paths to travel along given that we want to take the shortest route, but they can only be calculated if we know the entire structure of the network. So how do humans navigate networks with limited knowledge? Milgram's famous small-world experiment (Milgram, 1967) was the first to probe the size of the global social network by asking randomly chosen individuals to forward a letter via their first name acquaintances to a specified target individual. It demonstrated that there exist short paths between supposedly distant individuals. More surprisingly it also established that ordinary people without any special knowledge were good at finding them. So what is it about social networks that make them special? As Kleinberg (2000) pointed out it is their structure. On a random network short paths do exist but no one is able to find them due to the inherent randomness. With this in mind Peter Dodds and collaborators (Watts et al., 2002; Dodds et al., 2003) repeated Milgram's experiment using modern technology that allowed them to track the social search more thoroughly and pinpoint exactly which attribute was vital for a successful search. They found that while successful social search is conducted primarily through intermediate or weak strength ties, it depends on the incentive of individuals. This is directly in line with an earlier study performed by Marc Granovetter (Granovetter, 1973), who investigated how people get a job. Not surprisingly successful job hunting depends on the number of connections one has, however the type of connections is of even greater importance. As it turns out, it is not your closest friends that are of most use to you, as they are exposed to the same information as you. Paradoxically you have a higher chance of getting a new job through a weak tie, such as a distant acquaintance, because these weak ties act as bridges and can supply you with information you otherwise would not have received.

Ten years later, in May 2011, *Facebook* set in mind to test Milgram's finding and calculated the average distance between their then 721 million users⁷. Their result was not surprising; between any two individuals on earth, or at least amongst their users, the average distance was $\langle l_{FB} \rangle = 4.74$ (Backstrom et al., 2012), indicating that the diameter of our social network indeed is very small.

2.1.5 Clustering

In social networks it is very likely that the friend of your friend is also going to be your friend, this effect is called clustering or transitivity (Wasserman, 1994). In terms of network topology this effect manifests itself via an heightened number of triangles—sets of three nodes that are connected to each other. By counting the number of realized triangles versus the number of incomplete triangles in a network it is possible to define transitivity T as (Barrat and Weigt, 2000)

$$T = \frac{3 \cdot \text{number of triangles in the network}}{\text{number of connected triplets of nodes}}, \quad (2.7)$$

where a *connected triplet* denotes a single node that is linked to two other nodes forming an incomplete triangle⁸. The factor of three compensates for the fact that each triangle consists of three triplets, which in turn ensures that transitivity lies in the range $0 \leq T \leq 1$. In effect, transitivity measures the fraction of connected triples that are closed, meaning they have their third edge filled in to complete the triangle.

Watts and Strogatz (1998) proposed an alternative definition of clustering—local clustering—where they focus on individual nodes and formulate the clustering coefficient as

$$c_i = \frac{\text{number of triangles connected to node } i}{\text{number of triplets centered on node } i}. \quad (2.8)$$

This definition is used widely within the sociological literature where it is introduced as network density (Scott and Carrington, 2011). Global clustering for the entire network is then defined as the average of the local values

$$C = \frac{1}{n} \sum_i c_i. \quad (2.9)$$

By definition $0 \leq c_i \leq 1$ and $0 \leq C \leq 1$. The difference between calculating transitivity and clustering lies in the order of operations, Equation 2.7 estimates the

⁷This number has doubled since then, as of December 31st 2014, Facebook had 1.39 billion active monthly users (<http://investor.fb.com/releasedetail.cfm?ReleaseID=893395>).

⁸An easier graphical way to illustrate transitivity is $T = 3N_{\Delta}/N_{\wedge}$.

fraction of averages, while Equation 2.9 calculates the average of fractions (Figure 2.2 c). Regardless of which definition we use, clustering (or transitivity) is remarkably high for social networks compared to networks where nodes are connected at random. According to Newman and Park (2003) high levels of clustering, along with degree correlations, are some of the properties that make social networks special.

2.2 Network models

A network model is a way to explain the fundamental organizing principles of a system and to make predictions about future behavior. Constructing a network is nothing more than following a schematic that specifies the wiring diagram, denoting how to connect the various entities. This section will go through the most important networks models and focus on how successive discoveries about network properties have shaped our understanding of networked systems.

The simplest form of a network is given by a lattice (Figure 2.3 a). While its usefulness in modeling physical systems cannot be overestimated the general structure of a lattice model does not resemble any real-world network. This is because networks representing real-world systems are not regular, they are in fact objects, where order coexists with disorder.

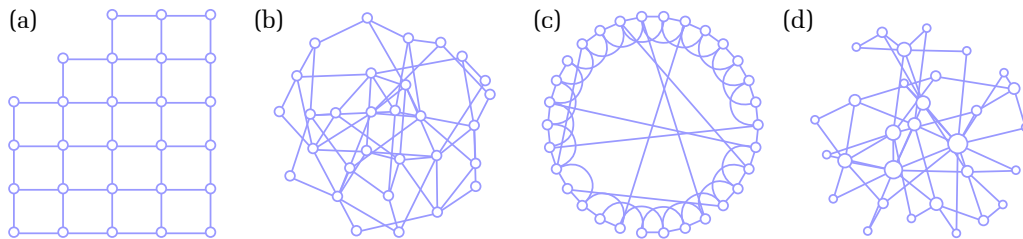


Figure 2.3: Network models. (a) Two-dimensional lattice model, where each node is connected to its four nearest neighbors. (b) Random network model proposed by Erdős and Rényi (1959), with $n = 30$ nodes and probability $p = 0.15$ of connecting pairs of nodes. (c) Small-world network by Watts and Strogatz (1998) with $n = 30$ nodes, $k = 4$ nearest neighbors, and rewiring probability $p = 0.1$ (d) Scale-free network of $n = 30$ nodes, constructed by the Barabási-Albert model (Barabási et al., 2000). Network is grown with the addition of one node and two edges per time-step. Nodes are sized according to their degree.

2.2.1 Random Networks

Proposed by two Hungarian mathematicians Erdős and Rényi (1959), random networks, or graphs as they are called in the mathematical literature, introduced

a new paradigm. The strength of random networks lies in their simplicity; defined entirely by the number of nodes n and the probability p of an edge existing between each pair of nodes (Figure 2.3 b). Due to this simple description many of their properties can be derived analytically. From their definition it immediately follows that the average number of edges in a random network is $\langle m \rangle = p n(n-1)/2$. Accordingly, the average degree of a node is $\langle k \rangle = n p$, which leads to the degree distribution assuming a binomial form

$$P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}. \quad (2.10)$$

For large networks, i.e. large values of n , this expression can be rewritten into the mathematically simpler Poisson distribution

$$P(k) = \frac{\langle k \rangle^k}{k!} e^{-\langle k \rangle}. \quad (2.11)$$

Random graphs have been extensively studied and shown to possess many interesting properties, we will briefly list some of them here, for a more complete overview see Albert and Barabási (2002). In terms of size, if the graph is connected, such that $\langle k \rangle \geq \ln(n)$, we can express the network diameter as

$$d = \frac{\ln(n)}{\ln(\langle k \rangle)}. \quad (2.12)$$

Thus the average number of steps between any two nodes scales as $l_{rand} \sim \ln(n)/\ln(\langle k \rangle)$. Effectively this implies that the “world” is small, making it possible to traverse the network with relatively few steps. With respect to clustering; if we pick a random node the probability of two of its neighbors being connected is

$$C_{rand} = p = \frac{\langle k \rangle}{n}. \quad (2.13)$$

As such, for large random networks ($n \rightarrow \infty$) the clustering coefficient becomes negligible.

Random networks are a good first approximation of real world systems, but as was pointed out by Kleinberg (2000) and Newman and Park (2003) real networks are found to be very unlike random graphs.

2.2.2 Small-world networks

In their seminal paper Watts and Strogatz (1998) tried to reconcile the terminology of random networks with the result from Stanley Milgram’s study (Milgram, 1967), which revealed that the distance between random individuals is very short. In fact, distances are so short that any pair of individuals, on average, are connected with just six steps, leading to the popular notion of “Six degrees of

separation” (Guare, 1990). Networks that possess this trait are called *small-world* networks. As it turns out, random networks display small-world properties because node-wise distances scale logarithmically with the number of individuals ($l_{rand} \sim \ln(n)/\ln(\langle k \rangle)$). The problem of random networks, however, lies in the clustering. Social networks display an unusually high number of triangles, i.e. situations where your friends are also friends, while random networks lack this property, because nodes, as the name implies, are linked at random. To solve this issue Watts and Strogatz viewed the problem from an entirely new perspective.

Their solution is based on a very simple idea; start from a regular lattice which has high values of clustering and mix in a little randomness to create the small-world effect (Figure 2.3 c). Beginning from a ring lattice where every node is connected to k individuals, they rewire links at random with probability p , enabling them to *tune* the network between order ($p = 0$) and disorder ($p = 1$); amidst the two the small-world property emerges. The small-world model captures most of the structural elements which we typically associate with a social network. We are all embedded in cliques with a very few close friends, but we also have distant weak ties with friends from childhood or with someone we have met while traveling—effectively making the world smaller.

There is, however, one problem with the small-world model by Watts and Strogatz (1998)—it is wrong! Its basic assumption about the structure of networks produces a Poissonian degree distribution similar in shape to random graphs (Equation 2.11). Effectively, this implies that nodes on average have similar numbers of connections and that networks have a typical size or *scale*. In stark contrast to studies that have discovered broad degree distributions in networks (Faloutsos et al., 1999; Barabási and Albert, 1999).

2.2.3 Scale-free networks

The small-world model was successful in incorporating short paths and high degrees of clustering, but failed in explaining why some nodes are considerably better connected than others. In 1999 two important papers by Faloutsos et al. (1999) and Barabási and Albert (1999) showed that *hubs* are an essential part of many real world networks, and that their degree distributions follow a power law

$$P(k) \sim k^{-\gamma}, \quad (2.14)$$

where γ varies from system to system but is always > 1 and typically $2 < \gamma < 3$ (Dorogovtsev and Mendes, 2002; Newman, 2003). Power laws are often called scale-free distributions, this is because they look the same on any scale. Note that for networks it is only their degree distributions that can be considered as scale-free. Mathematically this implies that the probability distribution satisfies

the criteria

$$p(\alpha x) = q(\alpha)p(x) \quad (2.15)$$

for any value of α (Newman, 2005). Plainly put, if we change the scale on which we study x by a factor of α , then Equation 2.15 tells us that its overall shape is unchanged except for a multiplicative constant $q(\alpha)$.

Power law distributions are not new, in a non-network context they have been known since the early nineteenth century. Vilfred Pareto, an Italian economist, was studying landownership in Italy and noticed a disparity of wealth where 80% of the land was owned by 20% of the population (Pareto, 1896). Later, American linguist George K. Zipf applied similar reasoning and formulated *Zipf's law*, accounting for the frequencies of words in the English language corpus (Zipf, 1949). In general, many properties such as city sizes, magnitude of earthquakes, and wars follow a power law (Newman, 2005).

Figure 2.4 illustrates the fundamental structural differences between random and scale-free networks. In a random network a typical node will have a degree of $k \sim \langle k \rangle$, as opposed to scale free networks where the lack of scale implies that a randomly chosen node can have a degree in the range of $k \sim \langle k \rangle \pm \infty^9$. Consequently, these highly connected nodes, denoted *hubs*, dominate the network topology (Figure 2.3 d).

For scale-free networks the distance between nodes scales as

$$l \sim \begin{cases} \text{const.} & \text{if } \gamma = 2, \\ \frac{\ln(\ln(n))}{\ln(\gamma - 1)} & \text{if } 2 < \gamma < 3, \\ \frac{\ln(n)}{\ln(\ln(n))} & \text{if } \gamma = 3, \\ \ln(n) & \text{if } \gamma > 3, \end{cases} \quad (2.16)$$

depending directly on γ (Cohen and Havlin, 2003; Bollobás and Riordan, 2004). For $\gamma = 2$ the degree of the biggest hub grows linearly as function of network size, this forces nodes into a hub and spoke configuration, where all nodes are a short, constant, distance from each other¹⁰. In the region where $2 < \gamma < 3$ the network distance grows as $\ln(\ln(n))$, thus comparably slower than $\ln(n)$ for random networks. This is denoted as the *ultra small-world* regime, where

⁹The degree of a node is of course limited to positive integer values $k = 1, k = 2, k = 3$, etc. $\langle k \rangle \pm \infty$ mainly indicates that the mean value is not descriptive of the system.

¹⁰Networks can also exist with $\gamma < 2$, this is called the anomalous regime. Here the largest hub grows faster than the size of the network. Thus for large values of n the biggest hub will have more connection than total number of nodes in the network, as $k_{max} \sim n^{\frac{1}{\gamma-1}}$ (Newman, 2003). This is only possible if self-loops and/or multiple links can connect the same pair of nodes. One therefore has to treat networks in this regime with great caution.

hubs connect many low degree nodes, radically shrinking the distance between them. For $\gamma > 3$ we observe the previously discussed small world networks (see subsection 2.2.2) where distance grows as $\ln(n)$. In between, at the critical point $\gamma = 3$, there is a mix of effects where the term $\ln(n)$ appears but is corrected by a double logarithmic correction. The slight difference between small-world and ultra small-world regimes may not seem like much, but if we consider a social network consisting of $n = 7\,000\,000\,000$ individuals, the average separation distance between pairs will be of the order $\ln(n) \approx 22$ (small-world), in contrast to $\ln(\ln(n)) \approx 3$ (ultra small-world). Demonstrating that hubs substantially reduce distances between nodes.

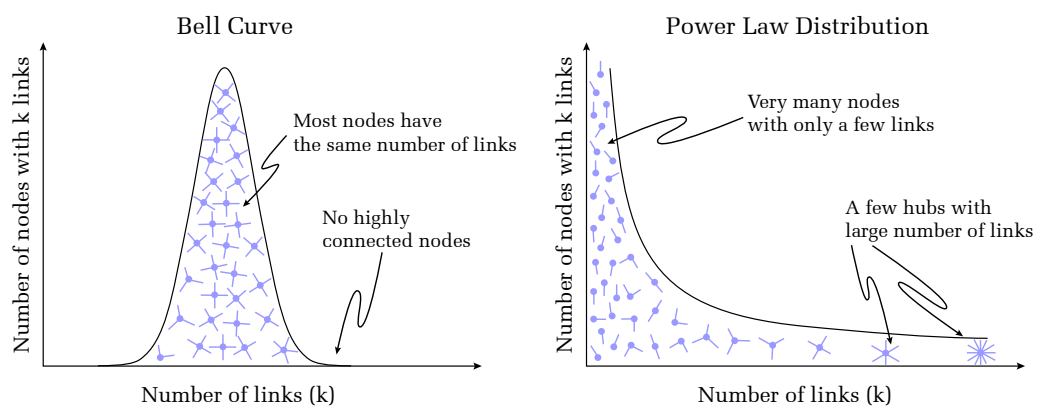


Figure 2.4: Network degree distributions $P(k)$, illustrating differences between a bell shaped Poisson degree distribution and a power law distribution. For random networks most nodes will have similar number of connections. In contrast to scale-free networks where a majority of nodes have very few links while few hubs are greatly connected. Redrawn from Barabasi (2002).

2.3 Network resilience

Networks rely on their connectivity to function, i.e. it is vital that paths exist between pairs of nodes. Removing nodes will increase the typical path length until a critical point where the network breaks up into separate components and communication becomes impossible. Thus, an important property of networks is their resilience towards node failure. Understanding how networks respond to node removal is important in many fields, in particular epidemiology; by removing (vaccinating) carefully chosen individuals we can split the network into separate components and stop the spread of diseases. As we, however, can remove nodes in a variety of ways, different networks will display varying degrees of resilience, but, as we will ultimately see, resilience is closely related to the degree distribution of a network.

Albert et al. (2000) were the first to investigate the resilience of scale-free networks with respect to node removal. They focused on two schemes of removing nodes, one removes nodes at random, while the second removes nodes according to their degree, removing high degree nodes first. Applying the schemes on representative samples of the Internet and of the World Wide Web, they looked at how the average distance between nodes was affected as they removed increasingly larger number of nodes. Figure 2.5 shows their results; both networks are relatively unaffected by random removal of nodes, while targeting high degree nodes has destructive consequences. In the context of scale-free networks we can intuitively understand this from the degree distribution. Because a majority of nodes have very few links, they will not lie on any vital paths, thus removing them will not affect the overall network. On the other hand, high degree nodes act as hubs for the network, i.e. many shortest paths travel through them, hence removing them will increase the overall path length. A great example is the worldwide airline network (Guimera et al., 2005), if we remove the most central airports we can still travel from city *A* to city *B*, but we would on average need to take more flights. The findings by Albert et al. (2000) suggest that scale-free networks are highly resilient towards node failures, but vulnerable against targeted attacks on its highest degree nodes. A comprehensive study of the resilience of complex networks has been performed by Holme et al. (2002), where the authors considered other strategies of node removal, and also looked into what happened if edges were removed instead of nodes.

2.4 Network growth

Up to now the discussed network models have tried to incorporate observed properties of real-world systems, such as clustering or length of paths, in an attempt to create networks that reproduce these patterns. The models, however, have not addressed the issue of how the networks in the first place acquired these properties. So in order to explain characteristic network features, such as highly skewed degree distributions, we shift our focus to models of network growth.

The Barabási-Albert model Not only did Barabási and Albert (1999) document that many real world networks follow power-laws, they also suggested a possible mechanism that explains how the phenomenon might arise. They argue that the scale-free nature of networks originates from two generic mechanisms: (1) most real networks are dynamic systems, which grow by the addition of nodes and edges, and (2) nodes do not link up at random, they follow certain preferences. The second mechanism is called *preferential attachment*. It states

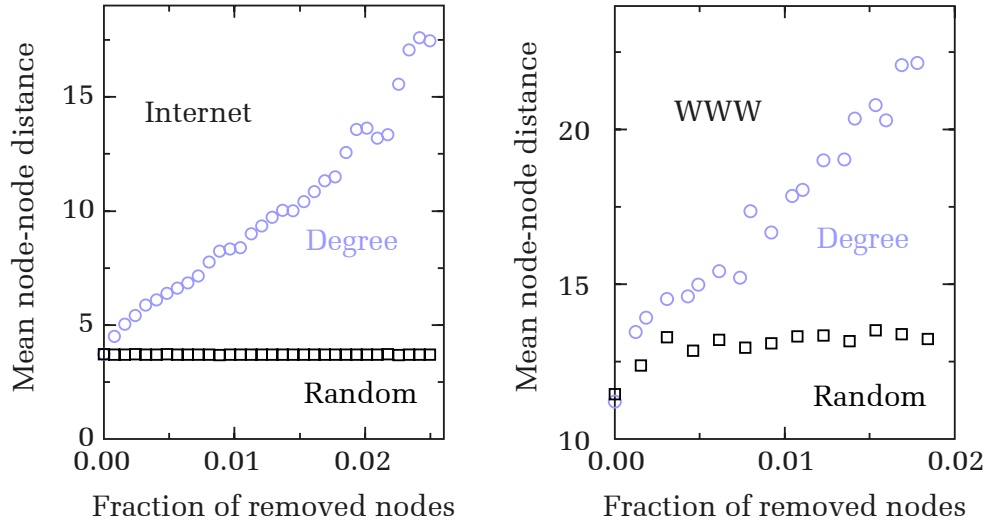


Figure 2.5: Network resilience, represented as the average node-to-node distance for increasingly larger fractions of removed nodes. Left panel shows the network of Internet routers, right panel show the network of web pages on the World Wide Web. Removing nodes in random order (\square) slightly increases the distance. Opposite, if nodes are removed in an organized fashion, largest degree first (\circ), distance increases dramatically, even for low fractions of node removal. Redrawn from Albert et al. (2000).

that new nodes link to preexisting nodes with probability proportional to their degree. In plain words this means that nodes favor to link up to already popular individuals. The concept of preferential attachment is not new, Simon (1955) originally used it in order to explain the unequal growth in wealth, called the “rich get richer” effect. 20 years later de Solla Price (1976) applied it in a network context to explain the observed skewness in the distribution of citations between scientific papers (de Solla Price, 1965). Suggesting that highly cited papers acquire citations at higher rates than less cited publications. Thus the model should arguably be called the Simon or the Prince model, but as the name, *Barabási-Albert model*, during the past 15 years has gained considerable popularity in the scientific discourse, we shall stick with it.

The Barabási-Albert model consists of three elements. (1) Start from an small initial configuration of nodes (n_0). (2) At every time-step add a new node with $m < n_0$ edges, linking it to m already present nodes in the network. (3) The probability, Π , for a new node to link to vertex i depends on the connectivity k_i , such that $\Pi(k_i) = k_i / \sum_j k_j$. Due to preferential attachment, nodes with high degrees will acquire new connections faster, hence any initial differences between nodes will be amplified. The authors further demonstrated that steps (2) and (3) are necessary in creating a power law, as absence of one fails to reproduce the desired behavior.

For the Barabási-Albert model it is possible to obtain an analytical expression for the power-law exponent, γ , which ultimately hints at the underlying topology of the scale-free network, see Equation 2.16. The rate at which node i gains connections is

$$\frac{\partial k_i}{\partial t} = m\Pi(k_i) = m \frac{k_i}{\sum_j k_j}. \quad (2.17)$$

As $\sum_j k_j = 2mt$, Equation 2.17 reduces to

$$\frac{\partial k_i}{\partial t} = \frac{k_i}{2t}. \quad (2.18)$$

Solving for k_i yields

$$k_i(t) = m \left(\frac{t}{t_i} \right)^{1/2}, \quad (2.19)$$

where t_i is the time at which node i was added to the system. Thus, the probability that node i has lower connectivity than k , $P(k_i(t) < k)$, can be written as $P(t_i > m^2 t / k^2)$. Given that we add new nodes at fixed time intervals, $P(t_i) = 1/(n_0 + t)$, we can write

$$P\left(t_i > \frac{m^2 t}{k^2}\right) = 1 - \frac{m^2 t}{k^2(n_0 + t)}. \quad (2.20)$$

The degree distribution can be obtained from

$$P(k) = \frac{\partial P(k_i(t) < k)}{\partial k} = \frac{2m^2 t}{n_0 + t} \frac{1}{k^3}. \quad (2.21)$$

In the asymptotic limit $t \rightarrow \infty$ it scales as

$$P(k) \sim k^{-3}, \quad (2.22)$$

with the exponent being independent of the number of edges (m) we add per time step (Albert and Barabási, 2002).

The Barabási-Albert model correctly constructs a scale free network, but as it turns out it has certain shortcomings in capturing other aspects of network structure. Krapivsky and Redner (2001) note that it introduces correlations between nodes, such that nodes of similar degrees are connected, however, as was discussed in subsection 2.1.2 this feature is only true for certain types of networks. In addition, Krapivsky and Redner (2002) also noted that power-law degree distributions only arise for a linear attachment of nodes. Another unnatural feature, which the authors themselves noted, is that the clustering coefficient vanishes as $n \rightarrow \infty$ (Albert and Barabási, 2002). Moreover, the model produces a correlation between the age of a node and its degree, i.e. older nodes will experience a higher increase of connectivity at the expense of younger

nodes. Adamic and Huberman (2000) showed that this prediction is inconsistent with empirically observed structural properties of the World Wide Web. In fact, a great deal of empirical evidence indicates that first movers do not necessarily end up as the most important ones, good examples are: *AltaVista*, once the most popular web search engine, it quickly lost ground to a latecomer—*Google*. Or the popular online social network *MySpace* which was later beaten by *Facebook*.

Responding to criticism Barabási et al. modified their model by introducing a novelty term (Adamic and Huberman, 2000; Albert and Barabási, 2002), such that each new node connects to already preexisting nodes with probability

$$\Pi = \frac{\eta_i k_i}{\sum_j \eta_j k_j}, \quad (2.23)$$

where η_i is the fitness of node i . This ensures that even relatively young nodes with fewer edges can accumulate connections at higher rates than older nodes. Nonetheless, it is an open question how to sample η_i , and from which distribution.

Finally, it has been shown that many other mechanisms can create power law distributions (Newman, 2005), thus preferential attachment might not be as important as previously assumed.

Despite the criticism, the model by Barabási and Albert was an important attempt in understanding the structure of networks, it represents a first simplified version of the world and highlights the importance of capturing topological features of real world networks.

2.5 Community structure

Networks are not organized at random; due to mixing effects, edges will be inhomogeneously distributed both globally and locally forming tight clusters of nodes. This feature is called community structure (Girvan and Newman, 2002). It is not a surprising fact, as we know from common experience that people segregate into groups along the lines of family, interest, age, occupation, ethnicity, and so forth. Within sociology it is therefore widely assumed that people form tightly connected groups, having a high density of edges within the group, and a lower density of edges between groups (Scott and Carrington, 2011; Wasserman, 1994). Figure 2.6 illustrates such structures revealing three distinct clusters.

Communities are not limited to only occur in social structures, but are observed in other networked systems from biology to the World Wide Web (Fortunato, 2010). Good examples are: clustering of web pages that deal with similar or related topics (Flake et al., 2002; Dourisboure et al., 2007), groups of proteins

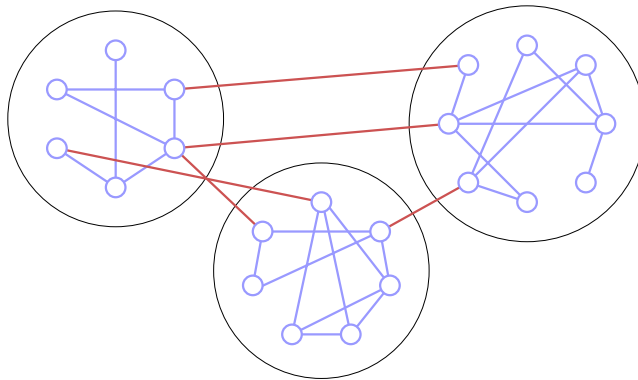


Figure 2.6: Community structure in a network, with three visible clusters denoted by black circles. Density of edges is higher within clusters, than outside. Edges between communities are colored red. Reprinted from Newman and Girvan (2004).

that have specific functions within our cells (Rives and Galitski, 2003; Jonsson et al., 2006; Spirin and Mirny, 2003; Chen and Yuan, 2006), or compartments within food webs (Pimm, 1979; Krause et al., 2003). The idea of identifying distinct groups of nodes is not new, Rice (1927) looked for clusters of people based on voting patterns, and Homans (1951) suggested a technique of identifying social groups by manipulating the rows and columns of adjacency matrices. By the 1950's and 60's computers had grown so large that electrical engineers faced the problem of partitioning electrical circuits into boards, dividing components into clusters by minimizing the number of cross connections (Kernighan and Lin, 1970).

Networks can display different levels of organization at different scales. This sometimes produces hierarchically nested structures with nodes being organized into communities which have smaller communities nested inside, which may again contain smaller communities, and so on. We wish to detect all these communities, thus we cannot focus on one specific community size, but need algorithms that can effectively infer these structures for us across multiple scales. One of the first algorithms to appear from within the network science community and one of the most popular algorithms nowadays, was proposed by Girvan and Newman (2002). In their seminal paper they suggest to identify communities using betweenness centrality. Edges with high betweenness act as connectors between otherwise separate clusters (Figure 2.6, red ties), removing them we can uncover the underlying community structure. The general idea of the algorithm is to (1) calculate the betweenness of all edges, (2) pick the edge with the highest value and remove it, (3) recalculate betweenness for the remaining edges, and (4) repeat from step 2. This procedure incrementally removes edges and builds a dendrogram of components, where at the lowest level, nodes are separated into individual communities. Run rampant this process will not find any meaningful partitions, but the question is when to stop it? To answer this, Newman and Girvan introduced a quantity called *modularity*, that measures how well a network splits into clusters. In their paper they argue that a good split is not necessary when there are few edges

between two components, but when there are fewer edges than expected. Thus modularity can be defined as:

$$Q = \begin{array}{l} \text{fraction of edges within communities} \\ - \text{expected fraction of such edges.} \end{array} \quad (2.24)$$

If the number of within-community edges is no better than what we can expect at random then $Q = 0$. Opposite, values of $Q = 1$ indicate good community division. In mathematical terms we can write Q as

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - P_{ij}) \delta(c_i, c_j), \quad (2.25)$$

where the sum runs over all pairs of nodes, m is the total number of edges in the graph, A is the adjacency matrix, and P represents the expected number of edges in the null model. The Kronecker δ takes the value 1 if nodes i and j are in the same community ($c_i = c_j$), and 0 otherwise. According Newman and Girvan (2004) very high values of Q are rare, for real networks high modularity values will lie in the range between 0.3 to 0.7.

The betweenness algorithm progressively removes ties and calculates Q for each partition; this is repeated until the process reaches the node level where no edges remain. The maximum value of Q then reveals the best place to stop the process and hence the optimal partition. In practice this algorithm works very well, but it suffers from one disadvantage: computation of all shortest paths for each iterative step is very expensive. Even for highly sparse graphs the complexity of the most optimized version of the algorithm scales as $\mathcal{O}(n^3)$, limiting its applicability to relatively small networks with $10^4 - 10^5$ nodes.

The goal of the betweenness algorithm is to find partitions with very high values of modularity. As it turns out, calculating betweenness is computationally costly, but why not just optimize Q directly? Newman and collaborators realized this possibility in two later papers (Newman, 2004b; Clauset et al., 2004), where they maximized Q using a greedy optimization scheme.

Today there are many strategies of optimizing modularity, we will not delve into them here, instead we focus on the choice of null model term, P . In principle this choice is arbitrary. We can, for instance, choose a network that has the same number of edges as the original graph, but demand that edges are placed at random. This will give us a random network with a Poissonian degree distribution, however, as was discussed in Section 2.2 this is not a good descriptor of real-world networks. A preferable null model should mimic the degree landscape of the original graph. Using the *configuration model*¹¹ (Łuczak,

¹¹ The configuration model is a generalization of random graphs to sample networks with preferred degree sequences. Start from n disconnected nodes and assign each node a number of

1989; Molloy and Reed, 1995) it is possible to calculate the probability that nodes i and j with degrees k_i and k_j are connected. To form a link between i and j we need to join two “stubs” (half-links). The probability to randomly join a stub attached to i with a stub incident on j is given by $k_j/(2m-1)$, excluding the stub already attached to i . As there are k_i chances of this occurring, the expected number of edges between nodes i and j is $P_{ij} = k_i k_j / (2m-1)$. In the limit of large m it reduces to $P_{ij} = k_i k_j / 2m$. Using this the expression of modularity yields:

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j). \quad (2.26)$$

Since only node pairs that are in the same cluster contribute to the sum, they can be grouped together and the sum can be rewritten to

$$Q = \sum_c \left[\frac{m_c}{m} - \left(\frac{d_c}{2m} \right)^2 \right], \quad (2.27)$$

where the summation now runs over all communities c , m_c is the number of within-community edges, and d_c is the sum of degrees of nodes in community c (Fortunato, 2010). Modularity maximization is widely applied in many algorithms, most notably in the *Louvain method* (Blondel et al., 2008).

Although no common definition has been agreed upon, communities are generally thought of as containing more internal than external connections. This was also the basic assumption of modularity. However, communities in networks often overlap, such that nodes can belong to several groups (Palla et al., 2005). In highly overlapping cases we even, counterintuitively, observe that communities have more external than internal connections. Thus, the general assumption of binary assignment of nodes breaks down, a new approach is needed. To solve this problem Ahn et al. (2010) introduced a new idea, instead of clustering nodes, why not cluster links? Their concept builds upon the observation that while nodes can belong to multiple groups, such as families, friends, and colleagues, links exist for one specific reason and within a single context. Rather than partition nodes into groups they consider a community to be a set of similar links. Since links that share a node are expected to be more similar than disconnected link-pairs Ahn et al. focus on such pairs and calculate the similarity between two links e_{iu} and e_{ju} as

$$S(e_{iu}, e_{ju}) = \frac{|n_+(i) \cap n_+(j)|}{|n_+(i) \cup n_+(j)|}, \quad (2.28)$$

“stubs” (half-edges) corresponding the the degree sequence of the original graph. To construct a randomized version of the graph, pair stubs at random. This yields a possibility of constructing self loops and multi-edges. The constructed graph will be an instance of the degree sequence of the original network.

where $n_+(i)$ denotes the set of neighbors of node i . Using hierarchical clustering, links are structured into a dendrogram, where each link is assigned to a single link community. By virtue of nodes having multiple links they can occupy multiple positions in the tree. As such, thresholding the tree nodes inherit all memberships of their links and can therefore belong to multiple communities. The dendrogram provides a rich hierarchy of structure where link communities can be extracted at multiple levels by cutting the tree. However, to obtain meaningful communities we need to determine the optimal place to cut the tree. For this purpose, the authors introduced an objective function, similar in purpose to modularity. The *partition density* measures the quality of a partition as

$$D = \frac{2}{m} \sum_c m_c \frac{m_c - (n_c - 1)}{(n_c - 2)(n_c - 1)}, \quad (2.29)$$

where n_c is the number of nodes in community c and where the contribution for a community is defined as zero if $n_c = 2$. The partition density measures how “*clique-ish*” versus “*tree-ish*” each community is. When every community is a fully connected clique $D = 1$, and when each community resembles a tree $D = 0$. If a partition is less dense than a tree, i.e. has disconnected components, then D can assume negative values.

Overall, knowing the community structure of networks is a valuable tool in understanding the topology and function of networked systems. Modularity and link clustering are just two of many methods that uncover this structure. In fact, community detection is the most prolific area of networks science producing an abundance of algorithms; in addition to modularity and link clustering some of the most popular algorithms are: clique percolation (Palla et al., 2005), Derényi et al. (2005), the Louvain method (Blondel et al., 2008), spectral methods (Newman, 2006), OSLOM (Lancichinetti et al., 2011), and InfoMap (Rosvall and Bergstrom, 2008).

2.6 Temporal networks

Human beings are not static, our behaviors, relations, and even our patterns of thinking change over time (Wrzus et al., 2013). Apart from certain growth models, where nodes in some sense are regarded as “*dead*” because once they arrive they never do anything new, we have so far viewed networks as static structures whose properties are fixed in time, but this could *not* be further from the truth. Intuitively we know that networks are constantly evolving, e.g. the number of web pages on the World Wide Web dynamically changes with some pages disappearing and new ones appearing. Like network topology, edges also display a temporal dimension, we know that our social ties are not continuously active, but only employed when we communicate. Networks, or

graphs, therefore have an inseparable temporal component where topology and edge activation profiles greatly affect the network behavior. In light of this, many of the traditional approaches to networks do not generalize to the temporal regime, and as we shall see, the temporal representation greatly complicates the description of networks. Nonetheless, the framework for studying temporal interactions is a valuable tool in understanding dynamic systems.

Many complex systems have an underlying temporal dependence and can therefore be described as temporal networks. Thus the study of temporal networks, in essence, is a very interdisciplinary field composed of disparate disciplines. This is reflected in the terminology, where one concept can easily have multiple names; to be consistent we will tie the formulation of temporal networks in with the previously discussed terminology in Section 2.1. For curious readers, a more thorough examination can be found in the excellent review by Holme and Saramäki (2012).

The rest of this section is structured as follows, first we highlight various real world systems that can benefit from being described with a temporal framework, followed by an overview of measures that quantify the properties of such networks. Lastly, we briefly discuss how to represent temporally evolving structures as static graphs.

2.6.1 Examples of temporal networks

Many complex systems might benefit from being represented as temporal networks. The following examples are, therefore, not at any level comprehensive, they merely emphasize that time is a vital component in complex networks.

Person-to-person interactions Social interactions are particularly suited to be described as temporal networks. Calls, texts, instant messages, and e-mails are all good example of dynamic processes where the spreading of information is temporally constricted. This representation can, among others, give us better insight into the dynamics of relationships (Holme, 2003), and how information spreads (Iribarren and Moro, 2009; Miritello et al., 2011).

A special type of human networks are proximity interactions, denoting who we are close to at what time. As we will see later in chapter 3, such data can be collected through various schemes and e.g. be applied in order to study the dynamics of diseases (Salathé et al., 2010; Stehlé et al., 2011a).

Cell biology Inside our cells, proteins are not expressed at all times. Since this is a waste of energy certain genes are only transcribed when they are needed. Incorporating time in a genetic regulatory networks Nelson et al. (2004) experimentally showed that oscillations appear, and Lewis (2003) demonstrated that

the formation of somites (precursors to the vertebral column) can only occur if constituent gene expression varies in time. Furthermore, considering the cell cycle in a temporal context with periodically and constitutively expressed proteins de Lichtenberg et al. (2005) experimentally demonstrated the presence of previously unknown modules and components.

Ecological systems The relationship between predator and prey is dynamic and changes with seasons, or over the course of an animals life cycle (Pahl-Wostl, 1995; de Ruiter et al., 2005). The evolutionary web itself evolves with species dying out and new ones arriving, although these processes occur over much longer periods of time.

Animals, or humans for that matter, are not stationary, we move around in space. Thus, population biology studies how proximity and mobility networks of animals change over time (Lusseau et al., 2003; Tantipathananandh et al., 2007). Much like humans, animal mobility is studied with respect to the spread of infectious diseases (Vernon and Keeling, 2009; Bajardi et al., 2011).

Neural networks Within our brains networks of neural connections represent another class of systems that can benefit from the temporal network approach. Over the course of a day, brain activity fluctuates with specific areas being active at different points in time depending on the mental task at hand (Dimitriadis et al., 2010). Valencia et al. (2008) showed that connectivity of functional regions changes in response to visual stimuli producing small-world characteristics. Bassett et al. (2011) demonstrated that the brain adapts its connectivity patterns in response to learning, reconfiguring edges between different regions of the brain, and Petri et al. (2014) showed that ingesting hallucinogens changes the brain's functional network.

Additional systems Other notable examples of networked systems that might benefit from the dynamic network approach are airline flight networks (Gautreau et al., 2009). Although flight connections are relatively stable each link has a dynamic weight (number of passengers) and a dynamically changing activation profile. In addition, the interrelations between banks, companies, or countries are all systems that might benefit from being modeled as temporal networks (Schweitzer et al., 2009).

In summary, complex systems are nontrivial combinations of evolving network structures and dynamical processes that take place on top of the networks. The question of whether a system should be modeled using the temporal framework ultimately relates to the timescales of the system (Gautreau et al., 2009). If the dynamical processes that unfold on the network are much faster than the

evolution of the underlying topology, then there is no need to model the system as a temporal network. As such, if networks change faster than the dynamics that occur on them, the formalism might be a good approach.

2.6.2 Properties of temporal networks

Temporal networks can be portrayed by three different representations (Figure 2.7). The *continuous* representation encodes interactions in a time-ordered sequence where each interaction can be viewed as instantaneous. This is typically used to represent communication events between individuals such as: e-mails, text messages, and calls, where we for the latter assume that the duration of a call is not important. The *discrete* representation divides time into windows and aggregates interactions that occur within each bin into a network. This approach is usually used to describe prolonged interactions, and can e.g. be used to characterize physical proximity of individuals. Aggregating interactions across all time yields the *static* representation, here all temporal information is lost. Regardless of which representation we use (continuous or discrete), including time in the network picture will often result in new and unforeseen behaviors, thus the statistical measures that we previously applied in describing static networks need revising.

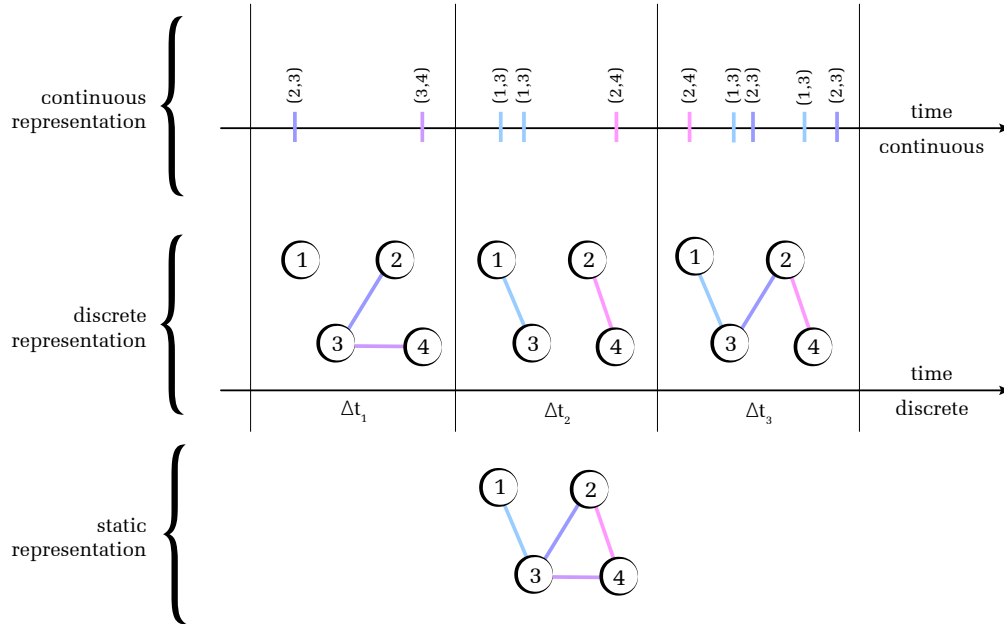


Figure 2.7: Temporal network representations of interactions between a set of nodes. The *continuous* representation consists of a time-ordered sequence of interactions. The *discrete* representation can be derived by segmenting time into equally sized windows and aggregating interactions that fall within. Aggregating events over all time produces the *static* representation. Figure is inspired by Williams (2013).

Degree distribution The degree distribution denotes the number of connections per node and characterizes the topology of a network. In a temporal context the degree distribution is best studied over longer periods of time, since time-resolved data is usually very sparse. In the case of high-resolution datasets there will only be a handful of edges present at each time-step. Figure 2.8 shows the growth of a network, and its subsequent change in topology, illustrated for various windows of aggregation (Δt). If we let $\Delta t \rightarrow \infty$ we arrive at the static network regime. Frameworks for finding optimal (also called natural) time-scales, where the conversion of dynamic data into discrete snapshots yields the best result, have been studied by Sulo et al. (2010), Clauset and Eagle (2012), and Krings et al. (2012). In general, there is no single optimal timescale, it depends entirely on the system of study and the dynamical processes that unfold on it. In addition, some systems might even display multiple time-scales comprised of daily, monthly, and yearly patterns.

Aggregating interactions is a simple way of removing the temporal dimension and allows us to apply the statistical framework from static graphs, but it can discard important information. Nevertheless, it is useful when the topological properties are of greater interest than the temporal.

Time-respecting paths Paths that connect nodes were previously thought as being ever-present. Thus if there was a path from a to b and from b to c then there was also a path connecting a and c . In the temporal regime, however, we need to take the time ordering of links into account. For nodes a and c to be connected there needs to be a sequence of time-ordered link-activations, such that the link (a, b) has to appear before (b, c) (Kempe et al., 2000). Consequently this implies that temporal networks are not necessarily transitive, meaning that there can exist a time-respecting path from a to c , but not from c to a . Furthermore, because paths are temporal by nature it is not guaranteed that a path joining a and c at time t will exist later at time t' ; moreover, future paths might even follow entirely different routes. Figure 2.9 illustrates the above concepts; node A is connected to node D via two time-respecting paths either through C or B . On the other hand, as interactions are not transitive node D is unable to interact with node A .

Time-respecting paths define which nodes can potentially be reached from other nodes, as a result they give us a clue about the reachability within a graph. Thus, the *influence* of node i is defined as the set of nodes that can be reached via time-respecting paths from i . This, in particular, is a valuable measure in understanding disease spreading, as it denotes the number of nodes that can be infected by i . Holme (2005) defined the *reachability ratio* as the average fraction of nodes that can be influenced. Conversely, one can also define the *source set* of node i as the set of nodes that can reach i via time-respecting paths, capturing potential sources of infection (Riolo et al., 2001).

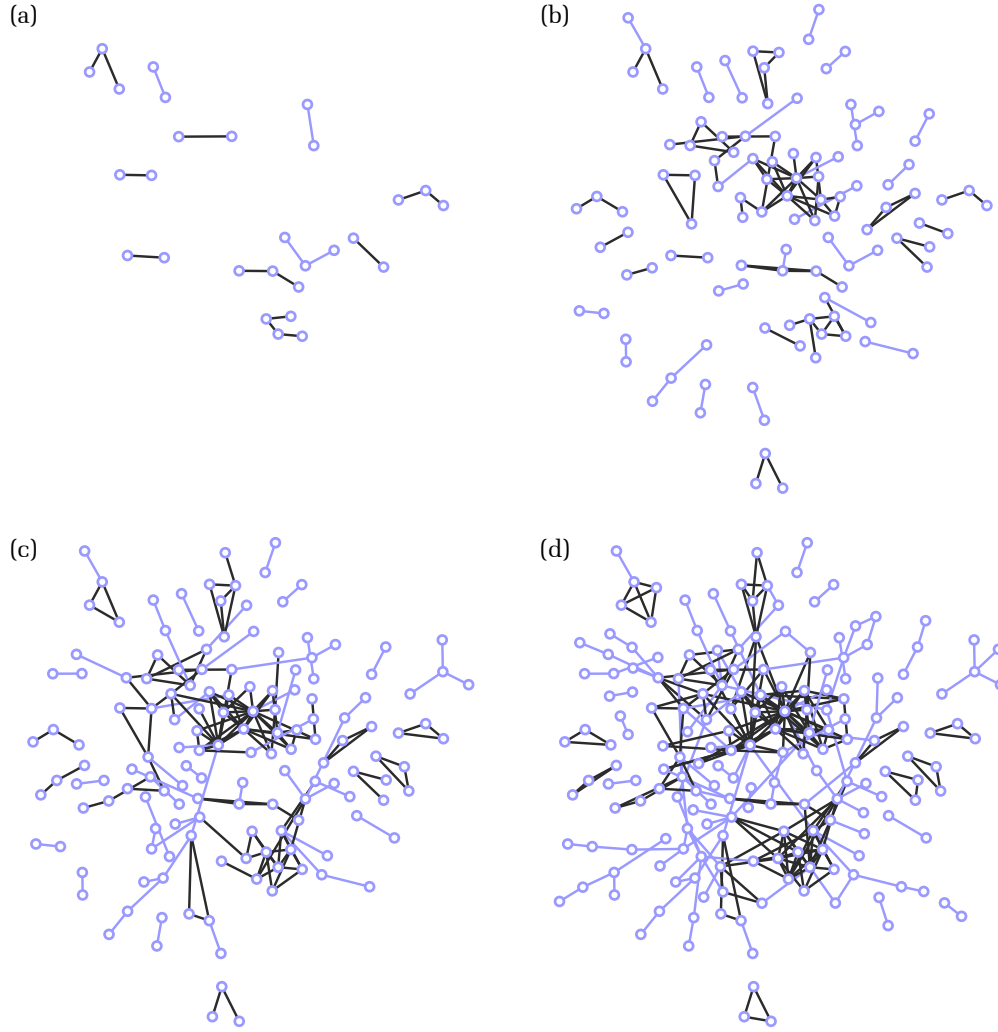


Figure 2.8: Network growth corresponding to aggregated call patterns within a subnetwork of individuals from a single postal code. Links that participate in triangles are colored black. Network is aggregated within intervals: **(a)** $\Delta t = 1$ day. **(b)** $\Delta t = 1$ week. **(c)** $\Delta t = 4$ weeks. **(d)** $\Delta t = 6$ months. Figure is adapted from Krings et al. (2012).

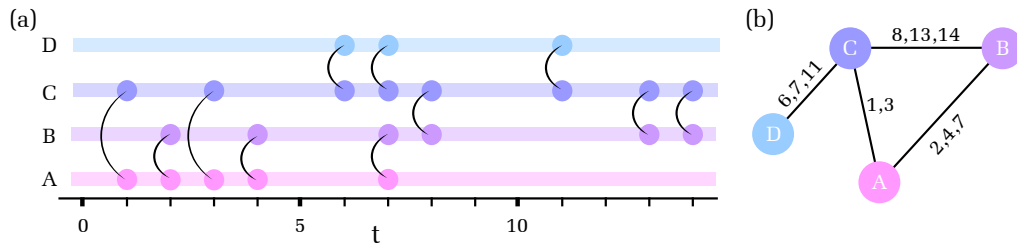


Figure 2.9: Time-respecting paths and the issue of transitivity. **(a)** Illustration of a temporal interaction sequence between four nodes. **(b)** Contact sequence from (a) visualized as a network, colored accordingly. Time of link activation are indicated on the edges. Adapted from Holme and Saramäki (2012).

Distances and fastest paths In static networks we defined distance as the number of ties one has to traverse in order to get from node i to j . Thus, it tells us something about the overall diameter of the network. Adding time to the formulation, we are no longer only interested in the length of a path, but also the time it takes to walk the path, a concept called *latency* or *temporal distance* (Pan and Saramäki, 2011). Introduced in the study of distributed computing, latency measures the age of i 's information about j (Lamport, 1978). Where it is assumed that nodes in contact update their information such that they possess the most recent. Thus average latency is a measure of how fast information propagates in a network. Empirically measuring the propagation of information has, however, proven difficult, as the generally accepted definition of latency relies on the existence of time-respecting paths. Various strategies that mitigate this issue are outlined in Pan and Saramäki (2011) and Karsai et al. (2011).

Centrality Centrality quantifies the importance of each node by e.g. counting how many shortest-paths pass through it or by its relative closeness to other nodes. Betweenness centrality is defined for static networks by Equation 2.6, this definition generalizes directly to the temporal regime by adding an explicit dependence on time (Tang et al., 2010):

$$b_u(t) \sim \sum_{i,j,i \neq j} \frac{\sigma_{i,j}(u,t)}{\sigma_{i,j}(t)}, \quad (2.30)$$

where $\sigma_{i,j}(u,t)$ is the number of shortest temporal paths from i to j in which node u has either received or relayed the information at time t , and the denominator, $\sigma_{i,j}(t)$, is the number of shortest temporal paths from i to j within a time-window ($t_{min} < t < t_{max}$). Other centrality measures such as closeness centrality, can also be generalized by adding an explicit dependence on time (Holme and Saramäki, 2012).

Inter-contact and Burstiness To supplement measures that characterize the overall structural and dynamical properties of temporal networks we can look at the constituent contributions from individual nodes and edges. Their associated activation sequences can tell us something about the correlation patterns between different entities, and might reveal important features about the timing and duration of time-respecting paths.

Human communication patterns have been shown to deviate from memoryless and random Poisson processes, instead they are concentrated in bursts (Barabasi, 2005; Oliveira and Barabási, 2005; Malmgren et al., 2009; Wu et al., 2010). This behavior manifests itself in the probability distribution of inter-contact times, $P(\tau)$, producing a broader than expected probability profile. However, due to certain inhomogeneities in the network structure, such as

heavy-tailed degree distributions, it is difficult to directly relate the inter-event distribution to the activation profile. It has therefore become common to display the rescaled distribution, $P(\tau/\tau^*)$, which compared to a Poisson distribution still displays a heavy tail (Candia et al., 2008; Miritello et al., 2011).

To quantify a signal Goh and Barabási (2008) introduced the coefficient of *burstiness*, defined as

$$B = \frac{\sigma_\tau - m_\tau}{\sigma_\tau + m_\tau}, \quad (2.31)$$

where σ_τ is the standard deviation, and m_τ the mean of the inter-contact pattern. Note, the definition is only meaningful when both mean and standard deviation exist. As this is always the case for real world signals, B is bounded to the interval $(-1, 1)$. $B = 1$ is the most bursty signal imaginable, $B = 0$ reflects a random process with Poissonian inter-contact times, and $B = -1$ corresponds to a completely periodic signal. Regular human heartbeat patterns have a burstiness coefficient around -0.75 , while communication patterns between humans are more bursty and lie in the range $0.2 < B < 0.7$ (Goh and Barabási, 2008).

Models of temporal networks Temporal network analysis is a relatively new field, and the number of models that explain the characteristics of temporal networks is rather limited. Few notable examples are: the model of social group dynamics by Zhao et al. (2011), and the models of temporal network structure by Jo et al. (2011) and Karsai et al. (2014). A more comprehensive overview is given in Holme and Saramäki (2012).

Including time in our representations of networks greatly complicates the mathematical description, but what do we gain in return? First, not all complex systems are suitable to be described by the formalism, it all depends on the structure and dynamics of the system. Secondly, bursty dynamics and time-respecting paths play a crucial role in spreading processes, be it the spread of cat-videos on social networks or the transmission of diseases, they influence these processes in new and unforeseen ways. Potential benefits of explaining such phenomena ranges from a deeper understanding of epidemics and how to prevent them, to the creation of more viral advertising campaigns. Last and most importantly, the paradigm of temporal networks represents a more realistic view of our world, enabling us to better understand its underlying driving forces. There are, however, still many open questions left, and according to Holme and Saramäki (2012): “*there is much room for improvement.*”

3

Measuring Networks

NETWORK science focuses on understanding complex natural phenomena by reducing them to the bare minimum—a collection of objects interconnected in some fashion. While a lot can be learned from theoretical approaches we need empirical data to understand real world systems. But what data should we collect and how should we go about collecting it?

There are enormous differences between collecting data in a classical sense and collecting data about humans. Atoms and cells do not mind if we observe, disturb, or prod them in order to uncover their fundamental secrets, nor do they protest when we change their environment. Humans, on the other hand, value the concept of privacy and we are not as keen at divulging our secrets. Because social systems are in a constant state of flux, one of the main pillars of the scientific method, *reproducibility*, does not apply, as it is virtually impossible to exactly replicate experiments in order to reproduce results. Thus, the challenges we face are colossal; to overcome them we need novel approaches.

In order to comprehend society we first need to understand the intricate web between people. To do so requires a detailed map of human interactions. Consequently, this chapter deals with ways of collecting, and measuring human behavior. As behavioral traces can be collected in a variety of ways, rather than giving a comprehensive overview of the methods of acquisition this chapter will focus on applying electronically collected traces as proxies for human life. Specifically we will discuss how recent technological advances have led to a better understanding of social systems and how new studies are currently

pushing the boundaries of quantitative social science even further. In this regard we will touch upon the dimensionality of data, can we view the world through a single lens or is that too simple of an approximation? Finally, this chapter concludes with an investigation of whether electronic traces can be applied to infer real physical interactions between individuals.

3.1 Proxy for human life

Recognizing genuine social relations is a central issue within multiple disciplines. When do connections happen? Where do they take place? And with whom is an individual connected? These questions are important when studying close-contact spreading of infectious diseases (Liljeros et al., 2001), or organizing teams of knowledge workers (Pentland, 2012). In spite of their importance, measuring social ties in the real world can be difficult.

For over a century social scientists have studied distinctive demographics using surveys, interviews, or by directly embedding observers within the environment of interest (Cooley, 1910; Moreno and Jennings, 1934). These methods capture rich sociological data, but are time demanding and labor intensive. Thus they are constrained to small sample sizes. With the rise of the information society our methods of capturing behavioral patterns have greatly improved. In fact our methods have improved so much that nowadays all our interactions with computers are stored in a database somewhere: every phone call, email, credit card transaction, shared picture, watched video, geographical check-in, and online comment is collected and stored somewhere. In addition, governments and cities record and collect comprehensive statistics about public spending, crime, and healthcare. Every aspect of our daily lives is being captured and stored in great detail, and the rate of information growth is expected to accelerate in the future (Lazer et al., 2009). These rich digital traces have the potential to revolutionize how we understand our society, and grant us the possibility to gain novel insights into human behavior and the underlying social fabric.

Collecting data has become relatively easy, but combining disparate datasets is still a challenge. Partially because of practical concerns such as noisy and incomplete data, but in part also due to complicated legal, ownership, and privacy issues of these highly sensitive traces. As a result most studies have concentrated on single samples of data such as emails (Kossinets and Watts, 2006), call detail records (Onnela et al., 2007), and online social networks (Lewis et al., 2008). These networks represent thousands to millions of individuals and display a wealth of interesting properties, but the rich information on interpersonal relations that previously was collected by human observers is now lost. Digital data traces offer a simplification of our world, and can can be

used to understand global features such as human mobility patterns (Gonzalez et al., 2008), but they fall short of describing the full complexity of human relations. This is because our interactions are expressed across a multitude of channels, we can talk face-to-face or over the phone, we can message each other using text messages, emails, or online social networks such as *Facebook* or *Twitter*. While we cannot communicate across all channels at once, we can effortlessly switch between them. Studying human relations using only one channel, therefore, does not yield a representative picture. The trade-off between rich ethnographic data and digital traces is complicated, and application of either data source depends on the system of interest. But is it possible to reconcile the two viewpoints and generate both rich and large-scale data?

We live in the era of mobile computing, where off-the-shelf mobile phones have more computational power than computers used by *NASA* when they first sent mankind to the moon. Nowadays smartphones can be modified in order to diagnose infectious diseases such as *AIDS* (Laksanasopin et al., 2015), or even be applied as portable brain scanners (Stopczynski et al., 2014b). The number of mobile phone subscriptions has grown rapidly over the previous 10 years (Figure 3.1). By the end of 2014 the number of active mobile phone subscriptions was expected to reach 6.915 billions, equivalent to roughly one device per every living person on Earth (ITU, 2014), making mobile phones the most pervasive piece of technology in human history. The ubiquity and functionality of mobile phones combined with our seemingly symbiotic relationship to them, makes them a perfect proxy for studying human behavior.

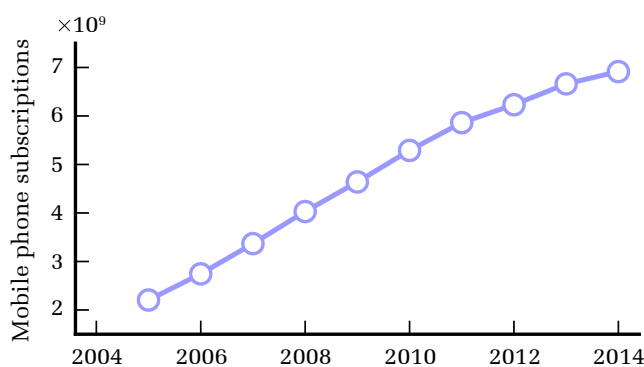


Figure 3.1: Growth of mobile technology. Figure shows the number of active mobile phone subscriptions. By the end of 2014 the number was expected to reach 6.915 billions. Data is obtained from the ITU (2014).

3.2 Data collection endeavors

Modern day smartphones come equipped with an arsenal of sensors and with enough computational power to rival personal computers. We are, therefore, not the only ones to have realized their potential, in fact there exists a large number of solutions for collecting behavioral data (Lane et al., 2010; Miller, 2012). Yet, few have attempted at collecting rich behavioral and longitudinal

data for relatively large populations. Examples are: the *Reality Mining* project where 100 mobile phones over a nine month period were applied as sensors to register social dynamics (Eagle and Pentland, 2006), and the *Social fMRI* study that collected the physical, online (*Facebook*), and credit card traces for 130 individuals over a 15 month period (Aharony et al., 2011). Lastly, the *Lausanne Data Collection Campaign* collected the daily traces of 170 volunteers living in the Lausanne area in Switzerland (Kiukkonen et al., 2010). Another interesting study that deserves a mention was performed by the *SocioPatterns* collaboration. They applied Radio Frequency Identification Devices (RFID) in order to study the dynamics of physical interactions (Cattuto et al., 2010). The downside of their approach is that individuals have to be outfitted with custom-made RFID tags and that interactions can only be studied in areas equipped with special radio beacons that communicate with the tags.

All the above mentioned studies have focused on relatively small and sparsely connected samples of individuals. For example the *Reality mining* project handed out 100 phones to students and employees at the *Massachusetts Institute of Technology*, yet only managed to capture 22 reciprocated friendships. While these studies have provided us with vital knowledge, they have been unable to provide us with fundamental insights into the dynamics of social systems. Fueled by the thirst for knowledge we want to go further and study the interactions between large numbers of densely connected individuals. Moreover, we want to study relations within a large number of different contexts, from workplaces to social settings, and from casual encounters to interactions between good friends. Thus, an unlikely but good place to start is at a university campus.

3.3 Copenhagen Networks Study

Aiming to push the boundaries of quantitative social science the *Copenhagen Networks Study* addresses the trade-off between rich ethnographic and longitudinal data by collecting information across multiple channels (Stopczynski et al., 2014a). Using state-of-the art smartphones as sociometers, the study collects data from a number of different sources regarding behavior, interactions, and demographics for approximately 1 000 densely connected students at a large European university. A series of questionnaires provide background information about socioeconomic, psychological, and health related factors for each participant. Data from *Facebook* produces a picture of the online persona, contributing with a view of the online social network and its dynamics. Sensors on the phones register and collect: geographic locations, telecommunication patterns, observable wireless networks, and face-to-face interactions between study participants. All this information is collected with a high temporal res-

olution down to the minute scale. In addition to the mentioned channels, an embedded anthropologist collects qualitative data about a subsample of the participants. All included this effectively makes the *Copenhagen Networks Study* the largest of its kind both in terms of covered information channels, and in number of participants.

These highly dynamic networks open up for new levels of observation. For example we are now able to study the diffusion of behavior, information, as well as infectious diseases. Some of these properties will only spread on specific types of links, e.g. we expect diseases such as influenza to spread via physical proximity, while information may diffuse across all types of links. As such we can study the differences and similarities between various spreading processes and find corresponding patterns. In addition to the richness of collected data, the setup of the *Copenhagen Networks Study* opens up for the possibility to run controlled experiments, addressing the notoriously complicated question of casual inference (Shalizi and Thomas, 2011).

The networks we collect are complex, entangled, and messy. Understanding them is a very difficult task, especially because the required knowledge to dissect their structures and dynamics is segregated across many scientific disciplines. Physicists and mathematicians have amazing analytical and computational skills, but they typically have very limited knowledge about individual behavior, social structures, and cultural norms. On the other hand, researchers from the social sciences, such as sociologists, psychologists and anthropologists, have spent a great deal of time thinking about such problems, but they lack the mathematical education. A deeper understanding of networks, therefore, requires collaboration across the sciences. The *Copenhagen Networks Study* is unique in this sense, as it is a collaboration between physicists, computer scientists, economists, anthropologists, psychologists, sociologists, philosophers, and public health researchers.

Summary of Paper I

Collecting rich behavioral traces from approximately 1 000 individuals along with their demographic and online information is a daunting task. In Paper I we present an overview of the *Copenhagen Networks Study* and describe in great detail the many considerations behind the experiment. More specifically we motivate our approach, outline the backend system, discuss important privacy issues, and suggest methods of returning data back to the participants.

Only preliminary results are presented, yet they clearly indicate the temporal aspects of human interactions (Figure 3.2) and show that human dynamics can unfold across multiple timescales. This has grave implications for the study of human relations as there is no simple way of aggregating data into a

stack of static networks. Instead this absence of a characteristic temporal scale accentuates the need for novel ways of perceiving dynamic data. Paper I further lists the statistical properties of each data source and e.g. illustrates that the distribution of personality traits within our sample is in alignment with previous studies performed on more diverse populations. Moreover, we demonstrate, as a proof of concept, how data from wireless access points (WiFi) can be applied to uncover physical interactions, effectively acting as a supplement to data collected from the proximity sensor. In cases where something goes wrong with the data collection or if users manually turn off their proximity sensor we can compensate by using this redundancy.

Each collected communication network reveals a specific aspect of human life and exhibits unique temporal properties, thus knowing only a single communication channel is not enough. Figure 3.3 shows the differences between (a) proximity networks, (b) call networks, (c) text networks, and (d) online networks. Providing evidence that social systems can look significantly different when viewed through different channels. Because we can, and do, communicate across multiple channels the scientific community needs to be aware of the limitations of using uni-modal data. In this regard, modeling the spread of information on a single communication channel will only uncover parts of the dynamics; while physical co-locations might be fully representative of the spread of certain diseases, they might not be entirely indicative of the spread of knowledge.

As the amount of collected data increases researchers will gain a better understanding of social systems and human nature in general, but we have to be careful not to draw these inferences from incomplete data sources. While the presented approach produces a dataset that is nearly complete in terms of communication between participants, it is clearly not the final answer. Since we only record data from a finite number of participants, our study population is a subset, and every network we analyze will be biased in some way (Kossinets, 2006). Nevertheless, our study represents a significant improvement in collecting, handling, and analyzing complex datasets.

3.4 Inferring relations

Our ability to collect and store data about networks has increased dramatically over the last few years, but how can we understand this data and what does it reveal? As networks are universal, we can be talking about friendships between individuals or physical connections between Internet routers, what properties should we measure and can we apply the same measures for all types of networks? As it turns out, it depends on the specific application in mind. The network features that reveal which individuals are most likely to be infected by a

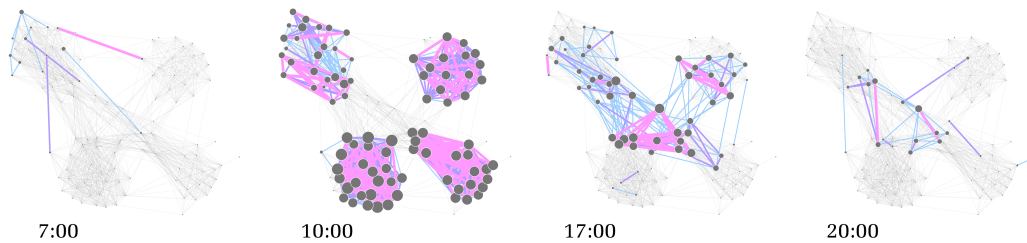


Figure 3.2: Dynamics of face-to-face interactions shown for a sample of the participants during four specific hours of a day. Students meet in the morning and attend classes within four different study lines (majors). Later during the evening they interact across majors. Edges are colored according to the frequency of observation, ranging from low (light blue) to high (pink). Nodes are linearly scaled according to degree.

disease may not be the same as the features that allow us to identify individuals most susceptible at adopting new technologies (Centola, 2010; Brockmann and Helbing, 2013). So, in order to understand networks we not only need to know how to characterize them but also whether the data we have collected can actually be applied for the purpose in mind. A lot of work has focused on the former issue, finding the correct characterization of networks properties (Costa et al., 2007), while little time has been devoted to understanding the underlying network data and whether we can make inferences about the social world from incomplete observations of events around us.

Applying data from *Flickr*, a popular online social network where users can share photos and interact, Crandall et al. (2010) siphoned through geographically tagged pictures and focused on events where pictures were taken by distinct individuals within the relative same geographical area at approximately the same time. From these coincidences they were successful in inferring online social ties. Suggesting that sparse data can be used in order to uncover parts of an online social network. Calabrese et al. (2011) went one step further and looked at the interplay between telephone calls, travel patterns, and geographic locations, in order to infer face-to-face meetings. Their results imply that physical meetings, to a certain degree, can be revealed by a combination of movement patterns where people travel to a prearranged location, and coordination behavior where individuals prior to a meeting exchange an increased number of calls in order to make the final preparations. While both studies uncover interesting properties about social structures, they ultimately do not reveal the entire truth because their data is too coarse-grained.

Using state of the art mobile phones as sensors Eagle and Pentland (2006) demonstrated a system for electronically sensing social interactions in their natural habitat. Combining observational data with self reports they further established the possibility of utilizing such proximity data in order to infer social

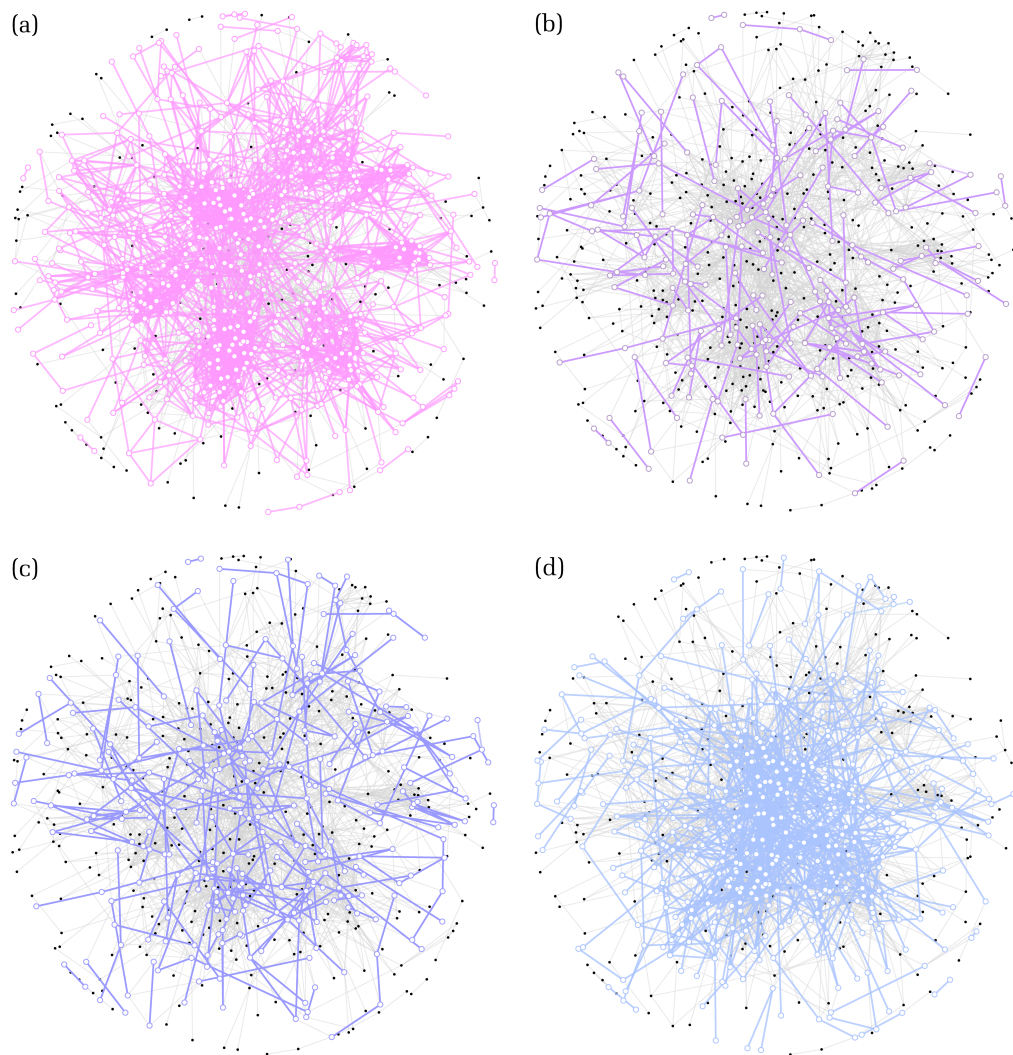


Figure 3.3: Perspectives of a social network, uncovered across four communication channels during one week. Focusing on internal interactions between participants in the *Copenhagen Networks Study*. (a) Top 5% most frequent face-to-face interactions, collected via the proximity probe. (b) Call network. (c) Text messaging network (SMS). (d) Online activity network from *Facebook*, denoting interactions where people either comment, share, or like each others posts. Due to privacy reasons private messages are not collected. Nodes are colored black and links gray if they are not present in a communication channel but appear in any of the others.

ties—accurately inferring 95% of reciprocated friendships (Eagle et al., 2009). Surprisingly, there are large differences between self-reported and electronic data, revealing that friends are more accurate in assessing their daily proximity patterns than pairs of individuals who do not consider each other as friends. This suggests that not all sensed ties are informative. As phones record proximity interactions within a sphere of radius 0 – 10 meters, they will undoubtedly infer false positive ties, sensing relations between individuals that are located in

different rooms, or on separate floors of a building. Furthermore, it is easy to think of examples where proximity does not correspond to social interaction, e.g. transient co-location in a dining hall. As such we have yet to understand to what extent electronic datasets may serve as a valid proxy for real life physical interactions.

Summary of Paper II

The use of mobile phones as scientific instruments is relatively new. Like any other instrument we first need to calibrate it, and understand what each measurement entails before we can apply the devices to study social systems. In Paper II we therefore focus on a sample of proximity data from the *Copenhagen Networks Study* and investigate how to identify real social ties and filter away noisy encounters. The proximity sensor collects data on the form (i, j, t, s) where each measurement implies that person j has been in proximity ($0 - 10m$) of person i at time t , where the devices have observed each other with signal strength s . Our investigation digs into the role of the received signal strength parameter and how this can be applied to distinguish between social and non-social interactions. Being present in a majority of other recorded proximity datasets, the signal strength parameter suggests a rough estimate of the distance between two devices. Intuition from physics tells us that the decay should be on the form $\sim 1/4\pi r^{-2}$. Nonetheless we perform an empirical study to confirm our reasoning. Taking pairs of devices we place them at distances $d = 0, 1, 2$, and 3 meters away from each other and measure the received signal strength. To check for signal variations phones are at each distance left to collect data for seven days. The empirical study confirms our general intuition, but it also uncovers interesting phenomena, such as bi-modal distributions. Yet, the study reveals that a simple signal filtering heuristic for a large majority of cases enables us to focus on interactions that occur within distances of $0 - 2$ meters. Such distances are in the social sciences denoted as a typical separation for interactions among close acquaintances (Hall, 1969).

To tie the findings from the empirical study with behavioral data from the participants, we look into what happens if we apply the heuristic and threshold *weak* (low signal strength) links. Behavioral data is divided into five minute wide temporal bins (Figure 3.4 a). Within each bin we remove observed edges if they have a signal strength below the threshold (Figure 3.4 b). We compare results to two reference models: a null model where we remove an equal number of ties, but where ties are chosen at random (Figure 3.4 c), and a control model where we remove the same amount of ties starting from the *strongest* ones first (Figure 3.4 d). We proceed to show that removing weak ties does not change the social networks structure greatly. Removing ties at random, however, has a

profound affect on network structure and significantly reduces clustering.

Disregarding weak interactions reduces the number of links in the network, but do we remove the correct links? The fact that clustering remains high in spite of removing a large fraction of links is a good sign. To investigate this question more directly we consider the probability that a removed link might reappear a short time later. Our results show that weak ties have a significantly lower probability of re-appearing in one of the subsequent time-steps than random ties. Regardless, the probability is not zero, because even weak links imply physical proximity. Similarly, if we remove a strong link, the probability of it reappearing is considerably higher than with a random tie. These results indicate that our proposed framework emphasizes real social connections while eliminating some noise.

Simply thresholding links based on signal strength is, however, not a perfect solution. Numerous scenarios exist where people are in close proximity of each other but are not friends, one obvious example is queuing. Each human interaction has a specific social context, and an understanding of the underlying social fabric is required to ascertain when a proximity link is a genuine social interaction. Paper II presents a first solution at the problem, a natural continuation would be to look into more complex features, such as co-arrival, co-departure and time spent together. In addition, telephone logs, online friendships and geographic locations are also factors which, coupled with proximity data, could give us a better insight into the dynamics of social interactions.

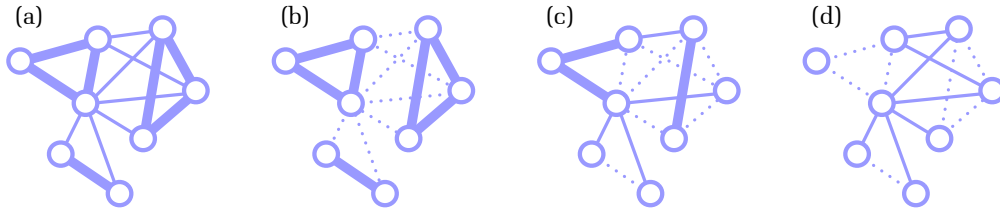


Figure 3.4: Heuristics of removing network ties in physical proximity networks. **(a)** Example of a typical proximity network observed within a time-window. The thickness of a link symbolizes the strength of the signal between two users. **(b)** Thresholded network, where links with signal strengths below a certain threshold are removed, dotted lines indicate the removed links. **(c)** Null model, where the same amount of links is removed but where ties are chosen at random. **(d)** Control network; an equal number of ties is removed, however, starting from the strongest first.

4

Understanding social systems

HUMANS have the potential to exhibit relatively random patterns of behavior, yet our lives are mainly dominated by routine. Viewed across different timescales these routines consist of daily practices from getting out of bed, eating breakfast, and commuting to work, to weekly, and longer yearly patterns such as holidays, where we e.g. prefer to spend religious holidays with our families. Although we rarely perceive our own actions as being random, for an outside observer they may seem highly unpredictable. All the choices we make are reflected in our mobility traces as well as in our social networks, which are in a constant state of flux. While we have seen impressive progress in understanding the basic laws that govern human motion (Gonzalez et al., 2008) little is known about the regularities of social systems.

This chapter deals with understanding human behavior, especially social dynamics. What role does randomness play in social life and to what degree can we describe and predict individual human actions? It is possible to study social systems from a wealth of different perspectives such as: network organization (Szell and Thurner, 2013), link formation (Wang et al., 2011), or contagion (Aral and Walker, 2012) among others. We choose to look at social systems from the perspective of sociality, where we focus on how individuals engage and participate in dynamic communities. The study of sociality is deeply rooted in the social science literature (Simmel, 1950; Goffman, 1967) and is a fundamental component of network science (Fortunato, 2010).

The chapter is structured as follows: first we discuss methods for uncovering

and extracting routine from behavioral traces. Then, employing proximity data collected by the *Copenhagen Networks Study* we investigate the social dynamics at microscopic levels of observation and describe a conceptually simple method of identifying evolving gatherings. From the temporal patterns of these gatherings, we infer dynamical communities, which in turn provide a simplification of the social system as a whole, resulting in a vocabulary for quantifying the complexity of social life.

4.1 Uncovering routine

Even though human life at times seems chaotic, it is imbued with regularities. Figure 4.1 shows an example of repeating patterns hidden within a dynamic social network. The figure shows that patterns of physical interactions, which we observe for the entire social system, are closely interrelated with interactions observed at later times, revealing correlations across days and weeks.

Certain patterns are easy to recognize, others are more subtle. For example the pattern of sleeping in is usually correlated with an activity in the previous evening, or to a specific situation, e.g. a weekend. These patterns can be difficult to observe, but become more apparent when put into a specific context such as temporal, spatial, or social. A traditional approach in modeling reappearing patterns is to apply Markov models. While they may work well for certain types of behaviors, they ultimately fall short in representing temporal patterns with multiple timescales (Eagle and Pentland, 2009). For such cases we need to adopt conceptually different methods. Eagle and Pentland instead propose a methodology that identifies repeated structures in the underlying behavior by breaking it down into principal components. Each component, or *eigenbehavior* as Eagle and Pentland call them, represents a characteristic set of features of an individual's behavior. A weighted sum of an individual's principal eigenbehaviors can then accurately reconstruct the behavior from each day. Eagle and Pentland further show that when weights are calculated halfway through a day, they can be applied to predict the behavior of the remainder of the day with 79% accuracy. Their method, however, can only be applied for short-term predictions as we need to recalculate and adjust the weights of the linear combination in order to accurately reconstruct behavior.

To explore long-term regularities in human dynamics Song et al. (2010) studied the patterns of human mobility by analyzing anonymized call detail records of mobile phone users. Every time a person makes or receives a call or text message the associated cell tower is registered. From this information they are able to reconstruct the trajectories of 50 000 individuals over a three month period. To measure the role of randomness and understand to what degree human mobility is predictable Song et al. apply entropy, an information

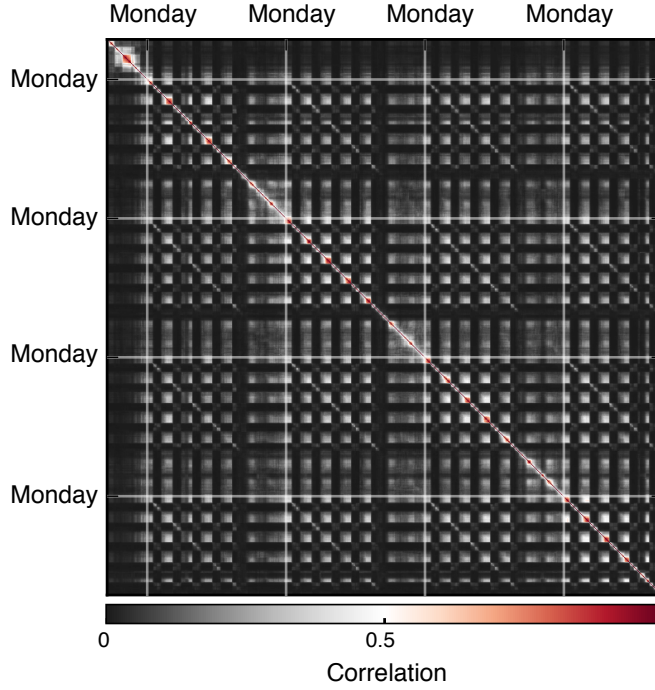


Figure 4.1: Regular patterns within a social network. Illustrating how physical co-locations are inherently routine based. Correlation is calculated as the overlap between the set of edges within two time-windows $\rho = |e_t \cap e_{t'}| / |e_t \cup e_{t'}|$. Interactions are shown for one month and aggregated within hourly bins.

theoretic measure, that characterizes the amount of uncertainty within a data stream. Given a sequence of states for an individual i we can define entropy in three ways. First we can think of entropy in a random sense, capturing the degree of uncertainty of a persons whereabouts if each location is visited with equal probability

$$S_i^{rand} = \log_2(N_i), \quad (4.1)$$

where N_i is the number of distinct locations visited by person i . Secondly, we can think of uncertainty in a temporally uncorrelated sense with entropy defined as

$$S_i^{unc} = - \sum_j^{N_i} p_j \log_2 p_j, \quad (4.2)$$

where p_j is the probability of observing state j . This version of entropy is usually denoted as Shannon entropy named after Claude Shannon (Shannon, 1948). Equation 4.2 captures the uncertainty of a person's location history by accounting for the frequency of states, but it does not take the order of visits into account. Thus it discards important information contained in the daily, weekly and monthly sequences of behavior. The *true* entropy of a person takes both frequency of states and the order in which they were visited into account. Let $T_i = [x_1, x_2, \dots, x_L]$ denote the sequence of states that user i has visited, *true* or *temporal entropy* is then defined as

$$S_i^{temp} = - \sum_{T'_i \subset T_i} p(T'_i) \log_2 [p(T'_i)], \quad (4.3)$$

where $p(T'_i)$ is the probability of finding a subsequence T'_i in the trajectory T_i . Clearly, $0 \leq S^{temp} \leq S^{unc} \leq S^{rand} \leq \infty$. The expression in Equation 4.3 is very difficult to evaluate. In practice one therefore resorts to estimating it numerically using Lempel-Ziv compression (Ziv and Lempel, 1978), which is known to rapidly converge to the real entropy of a time series (Gao et al., 2008). The Lempel-Ziv entropy estimate of a time series of length n is given by

$$S_{est}^{temp} = \left(\frac{1}{n} \sum_{j=1}^n \frac{\Lambda_j}{\log_2 n} \right)^{-1}, \quad (4.4)$$

where Λ_j is the length of the shortest substring starting at time step j which previously has not been observed. Further, Kontoyiannis et al. (1998) have proved that S_{est}^{temp} converges to S^{temp} when $n \rightarrow \infty$.

From the entropy of an individual (S_i) one can estimate the upper bound of predictability (Π_i) by solving a limiting case of Fano's inequality (Fano, 1961; Song et al., 2010; Jensen et al., 2010):

$$S_i = H(\Pi_i) + (1 - \Pi_i) \log_2(N_i - 1), \quad (4.5)$$

where

$$H(\Pi_i) = -\Pi_i \log_2(\Pi_i) - (1 - \Pi_i) \log_2(1 - \Pi_i). \quad (4.6)$$

Using the above methodology Song et al. (2010) estimated the upper limit of predictability in human mobility to be 93%, averaged across all individuals in their dataset. A limit which they found to be largely independent of the distance individuals cover on a daily basis.

Their method was a remarkable step forward in quantifying human dynamics, however, their results were obtained for meta-stable mobility patterns. We know that over the course of a human life our behaviors change. Caused either by transitioning from childhood to adulthood or by unfamiliar conditions such as emergencies (Bagrow et al., 2011). With this in mind Lu et al. (2012) looked into how the predictability of individuals changes after disasters. Collaborating with the largest mobile phone operator in Haiti Lu et al. analyzed the movements of 1.9 million mobile phone users prior to and after the devastating Haiti earthquake of January 12th 2010. Even though the earthquake greatly perturbed the lives of millions, their travel patterns, counterintuitively, still exhibited large degrees of predictability. Suggesting that our behaviors may be significantly more predictable than previously imagined.

4.2 Social groups

So far we have only considered regularities within our geospatial behavioral patterns, but our social life is also fundamentally based on routine. During

nights most of us sleep alone or with the same partner, in the morning we go to work and during the day we collaborate with colleagues; in the evening we hang out with the same gang of friends, and next day we more or less repeat the same patterns. These regularities in our social behavior are closely related to our geospatial patterns. For example interacting with colleagues is highly correlated with being at work, while nights are closely related to being at home. Thus complementary to mobility our social life should be equally predictable.

In order to understand the inherent patterns of social life, we first need to identify groups of interacting individuals. Luckily, within network science there is an entire discipline devoted to identifying structures within relations—it is called community detection. Yet, because human life is fundamentally dynamic we cannot apply the great body of knowledge about static networks, instead we need to view relations as temporal entities that can appear and disappear. While considerable amounts of work have been devoted in inferring communities in static networks, dynamic networks still pose a problem. Nevertheless, some progress has been made. Mucha et al. (2010) developed a generalized framework of quality functions that allowed them to extend the concept of modularity maximization for multislice and temporal networks. Modularity, however, has certain drawbacks as it experiences a resolution limit failing to identify clusters smaller than a certain scale (Fortunato and Barthélemy, 2007). Further, it has been shown that modularity exhibits degeneracies, lacking a global maximum and admitting an exponential number of distinct high-scoring solutions (Good et al., 2010).

In a more recent paper Gauvin et al. (2014) suggested the use of non-negative tensor factorization to detect dynamic communities. But, tensor factorization is a mathematically complicated optimization problem with two major drawbacks: (1) it does not have a unique global minimum and (2) we explicitly need to state the number of components it has to find. In their paper Gauvin et al. give examples of quality metrics that can be used to tune the number of components, but tensor methods are mainly viable in situations where we know the ground truth of the network, i.e. when we know how many clusters we should look for.

A more simple approach was proposed by Palla et al. (2007). They constructed an algorithm based on clique percolation that allowed them to track evolving communities. The percolation method builds communities from k -cliques, where each clique is a fully connected sub-graph of k nodes. Two k -cliques are adjacent if they share $k - 1$ nodes. A (k -clique) community is then defined as the union of all k -cliques that can be reached from each other through a series of adjacent k -cliques (Palla et al., 2005). Using this algorithm they successfully inferred dynamic communities in call and collaboration networks. However, the issue with clique percolation is the value of k , as it is not

clear what clique sizes one should choose and how this choice in general affects the quality of the inferred communities. Further, clique percolation is very sensitive to sparse data as it requires the presence of fully connected sub-graphs to function. Thus, for sparse networks, e.g. call networks (Krings et al., 2012), one needs to aggregate data into bins, but this inevitably leads us to the question of how to choose the appropriate bins width—a highly non-trivial question.

Each algorithm has its strengths and weaknesses. But the lack of benchmarks in the dynamic community detection field makes it difficult to pick the optimal one, as we cannot systematically compare and evaluate existing methods.

Instead, we choose to tackle the problem of identifying groups from an entirely new perspective. The remainder of this section is therefore devoted to describing our novel approach and exemplify its applications. First, we describe the basic principles behind the method and discuss how to sample dynamic networks with respect to time. Later we focus on how to track the evolution of social group and propose a scheme to identify re-appearing social structures. The following sections are part of the supplementary material of paper III, but are included here in order to present a coherent story. They will ultimately be tied in with the summary of paper III at the end of this chapter.

4.2.1 Network representations

Social networks are inherently dynamic, with nodes and links appearing and disappearing. When viewed at specific resolutions they reveal interesting properties (Figure 4.2). At daily, monthly, and yearly levels of aggregation we observe large densely connected networks that reveal the general structure of social systems and allow us to e.g. summarize their topological properties. Nonetheless, these large periods of aggregation obscure individual relations and make it nearly impossible, or at least very hard, to detect social groups. Hourly windows of aggregation supply us with a more local, yet still too aggregated view of social interactions. While a micro-level description disentangles the social web and directly uncovers group structures. Thus, when time slices are shorter than the group's turnover rate, we can directly and without ambiguity observe individual's group affiliations. Rather than identifying communities the challenge has shifted towards tracking group evolution.

4.2.2 Gatherings

It is possible to track the evolution of groups in many ways from simple matching schemes to more sophisticated machine learning techniques. Here we outline a simple method that applies hierarchical clustering to effectively track

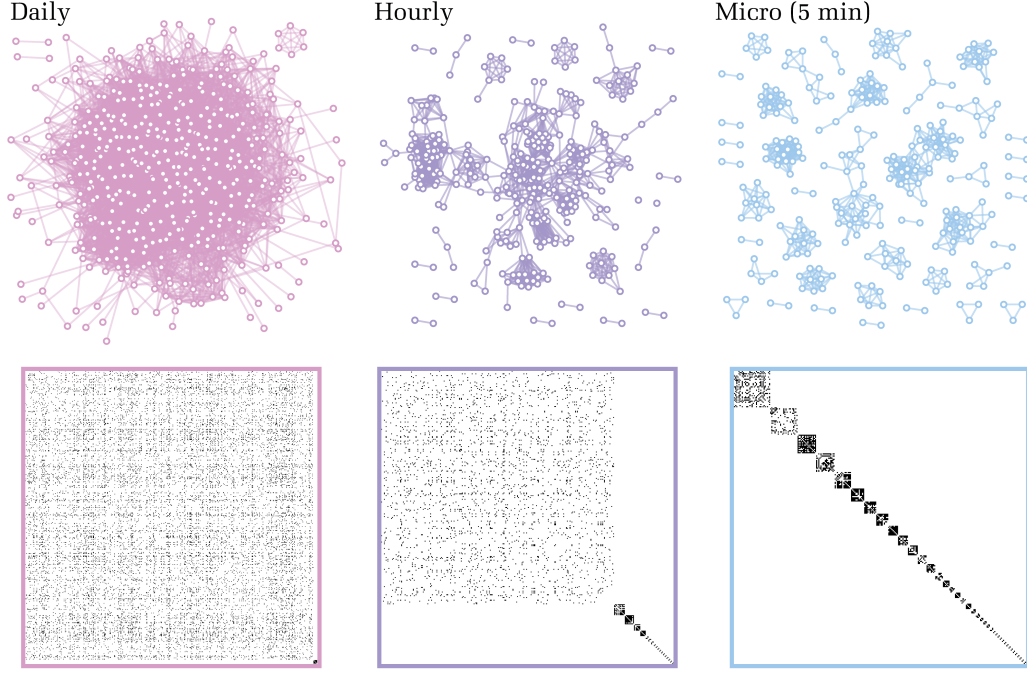


Figure 4.2: Physical proximity networks, constructed using daily (pink), 60-minute (purple), and 5-minute (blue) temporal windows. Below, corresponding adjacency matrices, colored appropriately, and sorted according to sizes of connected components.

group structures.

Dividing time into micro snapshots (Figure 4.2, right) we identify connected components, i.e. nodes that are in close physical proximity as social groups. Consequently, gatherings are defined as groups that are persistent across time. To infer gatherings we apply agglomerative hierarchical clustering—a widely used method that merges groups based on their similarity, S (Ward Jr, 1963). Initially, each group is assigned to its own cluster, then every iteration-step merges the most similar clusters according to the single linkage criteria ($\min(S(c_u, c_v))$; Gower and Ross, 1969). This merge criterion is strictly local and will agglomerate clusters into chains, a preferable effect when clustering temporal structures. The clustering procedure is repeated until all groups have been merged into a single cluster. This results in a dendrogram where each leaf is a group and where branches represent social gatherings. Similarity between groups is calculated using a modified version of Jaccard overlap

$$S(c_t, c_{t'}) = \frac{|c_t \cap c_{t'}|}{|c_t \cup c_{t'}|} f(\Delta t, \beta), \quad (4.7)$$

where c_t denotes the set of nodes present in group c at time t , $f(\Delta t, \beta)$ denotes coupling between slices, $\Delta t = t' - t$ denotes the temporal distance, and the β -parameter models decay. S is only defined for groups that share at least one node, as a result it is bounded between 0 and 1. The coupling

term, f , models memory (information) between temporal windows, where we assume that neighboring windows are related, but that this memory slowly disintegrates over time. The two most prominent forms of decay are exponential ($\exp(-\beta(\Delta t - 1))$), and power-law ($\Delta t^{-\beta}$). In the case of consecutive bins $\Delta t = 1$, hence $f(\Delta t = 1, \beta) = 1$ implying that there is zero decay (maximal memory) between neighboring slices. As Δt increases, S becomes negligible, and for *large* temporal distances we can completely disregard it and define the similarity as zero. In addition, as we are interested in the *future* evolution of groups, we require that $\Delta t > 0$. Combined, these two observations effectively reduce the computational overhead of the algorithm, since similarity only needs to be evaluated locally between groups observed within proximate time-slices that satisfy the criteria $t' > t$.

The outlined procedure constructs a dendrogram by repeatedly merging groups until all are assigned to a single cluster. This eventually forces highly dissimilar groups into single clusters. To extract meaningful social structures we need to partition the tree, but what quality function should we use? Modularity and link density have previously been applied (see Section 2.5) but they do not generalize well for dynamic processes, wherefore we need to construct our own. A preferable quality function compares the partitioned tree to some reference model, e.g. to the number of expected links. As we are, however, partitioning groups into gatherings we require a null model that comparatively mimics these structures, yet this is not straightforward as groups display non-trivial dynamics (Figure 4.3). Currently, the author is unaware of any reference model that might be applicable for such a situation.

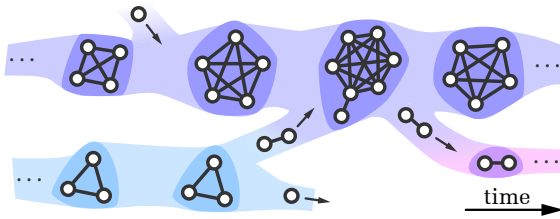


Figure 4.3: Gathering dynamics, with members flowing in and out of social contexts and with gatherings appearing and disappearing.

In lieu of constructing a reference model, we focus on the stability of inferred gatherings and argue that we should partition the dendrogram at the level where we uncover the most stable social structures. Each gathering is comprised of multiple groups, or slices, and can be thought of as the set $\{g_1, g_2, \dots, g_N\}$, where g_1 and g_N respectively denote the first (birth) and last (death) slice of the gathering. Gathering stability can then be defined in two ways. Local stability, calculated between consecutive slices as

$$\omega = \frac{\sum_{u=1}^{N-1} J(g_u, g_{u+1})}{N-1}, \quad (4.8)$$

and global stability, calculated between slices and the aggregated structure:

$$\Omega = \frac{\sum_{u=1}^N J(g_u, G)}{N}, \quad (4.9)$$

where J is the node-wise overlap between slices ($J = |u \cap v|/|u \cup v|$), and G is the aggregated structure ($G = g_1 \cup g_2 \cup \dots \cup g_N$). Both stability measures are defined as zero if gatherings are comprised of less than two groups. An analogous measure has previously been applied by Palla et al. (2007) in order to estimate the stationarity of communities. As each gathering has a specific stability, we are more interested in the average quantities of the entire system: $\langle \omega \rangle$ and $\langle \Omega \rangle$. If the dendrogram is cut too high, $S \rightarrow 1$, many gatherings will only consist of single groups, thus the stabilities will in that limit go towards zero. On the other hand if the tree is cut too low, $S \rightarrow 0$, $\langle \omega \rangle$ and $\langle \Omega \rangle$ will again go towards zero as many dissimilar groups will be clustered together. Somewhere in between the stability measures will achieve a maximum.

It is also worth noting that cutting the tree at values $S \leq 1/2$, may lead to unwanted side-effects. For example it is easy to imagine a scenario where two groups of equal size merge to form a new gathering or where a gathering splits into two equally sized parts. In such cases cutting the tree at values $S \leq 1/2$ will merge the gathering with both parts. We, however, find that a more desirable behavior is to declare the old gathering as dead and identify two new gatherings as born.

4.2.3 Dynamic communities

Each gathering contains information about its local appearance, so to gain a dynamical picture we need to match gatherings across time. A strict matching criteria is out of the question as gatherings have soft boundaries (Figure 4.4), with individuals participating unequally. An observations that is unlike the typical community detection assumption of binary assignment, see Fortunato (2010). We observe that some members participate for the total duration of a gathering, while others participate only briefly. The basic assumption of inferring groups as connected components will include noise, as people who coincidentally walk past a gathering might be included in it. One solution to this problem would be to prune such noisy interactions according to the heuristic presented in Paper II, however, as social interactions are not restricted to face-to-face meetings, this approach might cause more problems than it solves. Nevertheless, we need a method that lets us compare gatherings with varying levels of node participation.

Counting the number of times a node has been present in a gathering, we weight its participation relative to the total duration of the gathering. For exam-

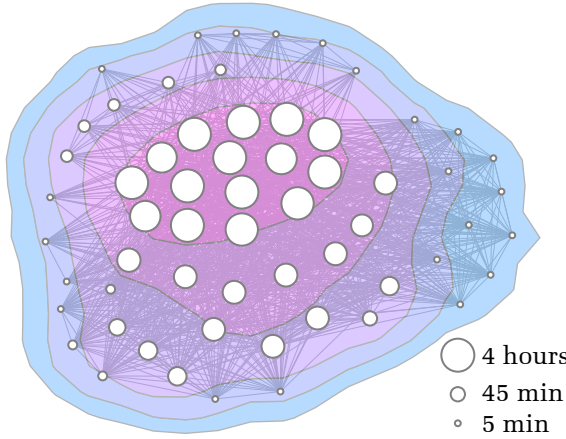


Figure 4.4: Gathering boundaries. Individual nodes participate in a non-homogeneous manner hence real world gatherings display soft boundaries, with nodes being organized into a stable core and a periphery. Node-sizes corresponds to participation.

ple if a node has participated in 80 bins within a gathering that has existed for 100 bins, we weight the node's participation relative to the gathering's lifetime and the node will get a weight of $80/100 = 0.8$. Because nodes no longer assume binary values, but may assume levels in the interval $0 \leq n_i \leq 1$ we calculate gathering similarity according to a continuous version of the Jaccard overlap

$$S(G_u, G_v) = \frac{\sum_{n=1}^N \min(G_u, G_v)}{\sum_{n=1}^N \max(G_u, G_v)}, \quad (4.10)$$

where G_u is a vector containing normalized node-wise participation values for gathering u , and N is the total number of nodes in $G_u \cup G_v$. The functions \max and \min act piecewise on the two vectors, and $S(G_u, G_v) = 0$ between two gatherings with zero node-overlap. Using this metric we look for underlying patterns in how nodes participate in gatherings. To uncover these structures we apply an agglomerated hierarchical clustering scheme with the average linkage criterion¹². Iteratively this method produces a dendrogram with gatherings as leaves. Partitioning the tree clusters similar gatherings together into communities. As each gathering appears at different points in time we can think of the communities as being dynamic.

A community consists of all nodes from its constituent gatherings, but it also inherits their individual participation vectors. Thus we need a method to construct a community participation profile from its gathering subcomponents. This can be done in many ways, we chose a version where we weight the participation vector of each gathering with its respective lifetime (τ_G), producing the community participation vector:

$$C = \frac{1}{\sum_{G \in C} \tau_G} \sum_{G \in C} \tau_G G. \quad (4.11)$$

¹²It is possible to apply other linkage criteria as well, such as complete or Ward-linkage. All three types of criteria yield comparatively similar results.

Facing the same challenge as before (section 4.2.2), we again have to estimate the optimal place to partition the tree. This time, however, we are in luck as we can solve the problem using standard machine learning tools. The *gap statistic* was originally proposed by Tibshirani et al. (2001) as an unsupervised method for estimating the number of clusters by finding latent structures within the data. Their method is based on calculating the change of error between partitions as the number of clusters is increased, compared to a similar change within an appropriate reference model. Given a total of g gatherings clustered into k communities, C_1, C_2, \dots, C_k , we calculate the within-cluster error, or dispersion as

$$W_k = \sum_{r=1}^k \frac{1}{2|C_r|} \sum_{u,v \in C_r} D_{uv}, \quad (4.12)$$

where $|C_r|$ is the cardinality¹³ of cluster r , and $D_{uv} = 1 - S(G_u, G_v)$ is the pairwise distance between gatherings u and v where S is defined in Equation 4.10. Thus W_k is the accumulated sum of within-cluster differences around the cluster mean, and the factor of two takes double counting into account. The idea behind Tibshirani et al. approach is to compare this difference to the expected within-cluster error generated by a reference model (W_{kb}), defining the gap statistic as

$$\text{Gap}(k) = \frac{1}{B} \left[\sum_{b=1}^B \log(W_{kb}) \right] - \log(W_k), \quad (4.13)$$

where we average over B reference datasets. The optimal number of clusters which we should partition our data into is the value of k for which $\log(W_k)$ falls furthers below the reference curve after we have taken the sampling distribution into account, i.e. it is the value of $k = k^*$ such that

$$k^* = \underset{k}{\operatorname{argmin}} \{k \mid \text{Gap}(k) \geq \text{Gap}(k+1) - \tilde{s}_{k+1}\}, \quad (4.14)$$

where $\tilde{s}_k = s_k \sqrt{1 + 1/B}$, and s_k is the standard deviation of $\log(W_{kb})$ over the B reference datasets.

In their paper Tibshirani et al. (2001) give examples of how to construct simple and effective reference models. They have two suggestions: (1) generate reference models by sampling features uniformly over the range of the observed values for each feature, or (2) generate features from a uniform distribution aligned with the principal components of the data. Given that we cluster gatherings, our features include nodes and their respective participation levels. Finding principal components is not really an applicable method as data at the node-level is one dimensional. Rather, we follow their other suggestion and

¹³Cardinality of a set is the number of elements in the set. For example the set $A = \{1, 2, 3, 5, 8\}$ contains 5 elements and therefore has cardinality 5. Hence the cardinality of a cluster is the number of gatherings that belong to it.

construct reference models by assigning random participation values sampled from a uniform distribution. In detail, we assign participation values uniformly sampled from the interval $(0, 1]$ to randomly chosen nodes. This is done in accordance with the size distribution of gatherings, such that our reference models reproduce the underlying structural aspects of social meetings.

The gap statistic has certain drawbacks; for example it is notoriously slow for large datasets as it needs to compute the within-cluster dispersion for all partitions from $k = 1$ to $k = N$, where N is the total number of data points. For such cases there exist computationally fast approximations of the gap statistic (Giancarlo et al., 2008).

In order to identify both gatherings and dynamic communities we applied hierarchical clustering and partitioned the corresponding dendrogram in the most optimal place, found by a quality metric. An interesting point, however, was made by Pons and Latapy (2011): when partitioning a dendrogram why constrain ourselves to a *straight* horizontal cut? As some parts can be higher or lower in the tree, a horizontal but not necessarily straight cut might reveal better quality partitions. This insight is especially intriguing and relevant for our approach as we partition the dendrogram according to average values of the quality metrics. In a worst case scenario it might turn out that while the cuts are optimal for the average values they are not optimal for any single gathering or community. Thus Pons and Latapy's idea is a good way of improving our proposed approach.

The past sections have been heavy with methods but sparse with results. To counteract, we will apply the outlined methods in hope of understanding and quantifying the behavior of individuals. The next section highlights the work performed in Paper III, and focuses on describing individuals from the aspect of their sociality.

Summary of Paper III

A wide range of applications from predicting the spread of epidemics, to city planning, and resource management can benefit from an ability to foresee human behavior. Where, when, how, and with whom? All are questions that try to probe a specific aspect of human life, but is human behavior predictable? By measuring the entropy of human travel patterns Song et al. (2010) demonstrated that we are less chaotic than we usually think, and quantified that human mobility can be correctly predicted with up to 93% accuracy¹⁴. While we have

¹⁴Even though I do not entirely believe in their results as their prediction problem deals with guessing the location of an individual in the next time-bin, rather than guessing the next location; their approach was a remarkable leap forward in quantifying human behavior. Yet,

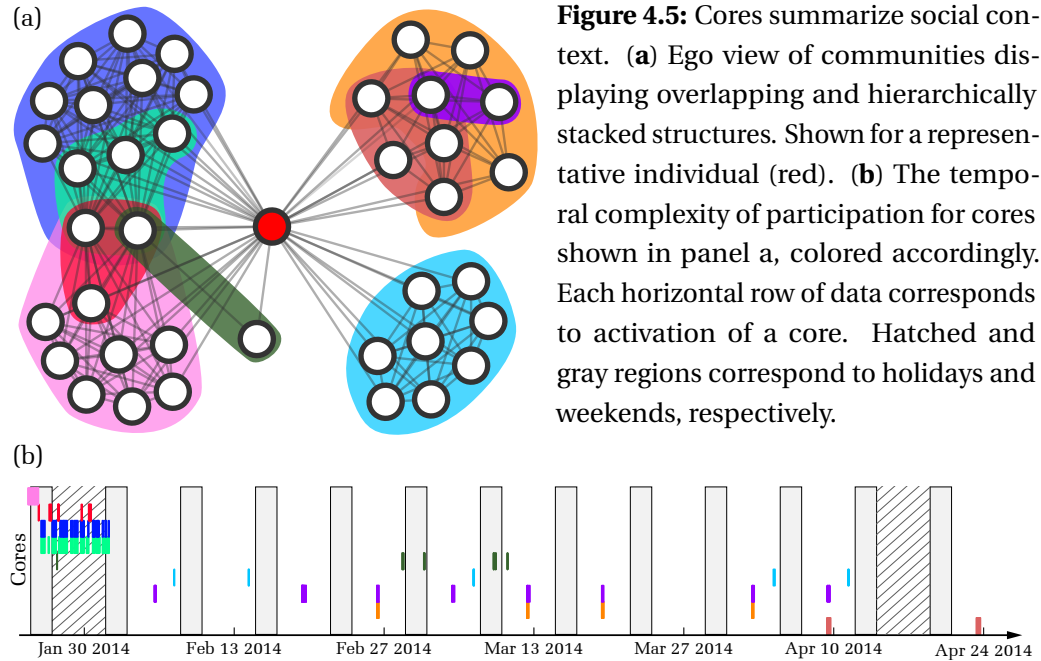
seen impressive progress in understanding the fundamental laws that govern human travel patterns and complex systems in general, little is known about the regularities governing social networks.

Based on a unique data-sample from the *Copenhagen Networks Study* we consider the social relations between approximately 1 000 densely connected individuals over a period of 5 months, representing roughly one semester. We show that high-resolution data untangles the intricate web of relations and allows us to observe social groups directly and without ambiguity—making traditional community detection algorithms redundant. A simple clustering scheme across time-slices then reveals the temporal development of each gathering, where dynamically evolving gatherings present a fundamentally new entity for quantitative study. While nodes can only be part of one gathering per time-step they can effortlessly switch affiliations between coinciding gatherings. It is due to this gradual exchange of nodes that community detection has proved difficult in previous settings. Unlike the conventional community detection assumption of binary assignment we observe that gatherings have soft boundaries with members coming and going, but organized via a stable core of individuals who are present throughout the entirety of the meetings. From repeated appearances of cores we then infer dynamic communities. Some communities are only observed once while others are more active, appearing, on average, multiple times per day. This results in a broadly distributed number of appearances. Dividing communities into *recreation* and *work* settings, depending on where they are observed, allows us to distinguish between schedule driven and social behavior. We observe a clear difference in how individuals engage and spend time with respect to varying social context; for example recreational meetings tend to be smaller but last considerably longer.

Representing a persons social interactions in terms of cores provides a powerful simplification of a dynamic social network (Figure 4.5 a). Instead of the constant and unwieldy flow of interactions that we observe in raw data, cores represent a set of states for quantifying social life. In analogy to geospatial behavior we can think of human social life as a temporal sequence of *social states* (Figure 4.5 b). Given this sequence, we can use information theory, specifically temporal entropy, to quantify the predictability of social life. Additionally, as we have access to detailed mobility patterns for each individual (Cuttone et al., 2014), we can compare the limits of predictability for both spatial and social aspects of human life. Surprisingly, we find that social interactions contain even higher levels of predictability than human mobility patterns, however, we also

we show in the supplementary material of paper III that it is possible to achieve arbitrary high values of prediction by narrowing the bin-size. For example, we can predict individuals with approximately 98% accuracy if we focus on 30-minute wide bins instead of the 60-minute bins applied by Song et al.

demonstrate that the overall levels of social and spatial predictability are not correlated.



In general, predictable geospatial behavior does not imply predictable social behavior, or vice versa. While the two types of behavior are not correlated over longer periods of time there is a subtle interplay between them as they both are closely related to daily and weekly schedules. We find that during the week, our social and spatial behaviors are entangled; we tend to meet the same people in the same places. However, during weekends this interrelation is reduced. While our mobility traces shows clear signs of exploratory behavior during weekends, our social interactions become simpler and more consolidated. We use this to propose a new kind of prediction—social prediction. It is a well known fact that individuals who share a social tie are predictive of each other (Crandall et al., 2010; Cho et al., 2011; De Domenico et al., 2013a). Nonetheless, friendships lack a temporal signature informing us exactly when individuals are predictive. Cores, on the other hand, provide such a temporal context, where an incomplete set of core members implies that the remaining members will arrive shortly. We demonstrate this concept for cores of size three. Given that two members are observed we calculate the probability of the remaining individual joining within one hour. We compare this to two reference models: (1) where we construct reference groups by randomly picking individuals and (2) where we use pairwise friendships to construct groups. To avoid testing on meetings that are driven by the academic schedule, we focus on weekends and weekday evenings and nights (6pm-8am). Furthermore, we test on a month of data that has not been applied for identifying cores. Our results show that cores greatly

outcompete both reference models. Which in turn illustrates that pairwise friendships are not enough to predict future interactions of people. Rather, it is the signal encoded within the context of the meeting that is important; the reason cores are able to predict the arrival of individuals is because social context requires all core members to be present.

The presented work is a first descriptive effort at quantifying the rich patterns encoded within social systems and suggests a new framework for describing human behavior. It is generally believed that incorporating a temporal dimension complicates our description of networks. In Paper III, however, we find that community detection becomes considerably simpler as we can directly observe communities on short timescales. Utilizing this, we identify gatherings that consist of stable cores of individuals. These cores then provide a powerful new description of social behavior which, among others, can be applied to characterize the predictability of social life. The paper does not claim that the subjects in the *Copenhagen Networks Study* are a fully representative sample of society. Nonetheless, many people exhibit regularities in their daily life, from getting up in the morning, eating breakfast, traveling to work, and returning home in the evening. As such, routine is not an exclusive trait of our population. In fact we can imagine that students with large amounts of spare time and constantly evolving social networks are even more unpredictable than a typical human being.

5

Summary

NETWORKS are a fundamental part of life. Understanding them is vital if we are to gain insights into the complex working of nature, society, and humans. As networks are in a constant state of evolution, we cannot approach them with the traditional paradigm of static graphs. Instead we need new approaches that reconcile the worldview of static graphs with the complexity of dynamic networks. Yet, before we can start to unravel the mysteries of dynamic networks, we first need maps, not just any type of maps but highly detailed and accurate maps.

In this dissertation we have described how to collect, measure, and analyze complex social networks. We have emphasized that human life cannot be viewed through any single network layer as we interact over a wealth of communication channels. As such information can diffuse through: face-to-face interactions, calls, text messages, *tweets*, emails, interactions on online fora, and even television and radio broadcasts. Further, life cannot be perceived at any single temporal scale as all our networks are ever-changing. Nevertheless, each time-scale can be used to view traditional problems from new perspectives.

Network data usually comes in unstructured and messy batches. Often we do not even know whether the data is representative of the object we want to study. Exploring the question of sociality, we demonstrated how to infer social relations from noisy proximity traces. Although our solution is not perfect it serves as a first iteration and illustrates that a simple mechanism is surprisingly efficient in emphasizing social encounters.

Collecting network data with high resolution gives us a new outlook at the microscopic levels of social systems. From this data we show that one can easily and without ambiguity identify dynamic communities, making traditional community detection heuristics redundant. Contrary to the usual assumption of binary assignment of nodes into clusters we show that people participate in non-homogeneous ways, forming stable cores within each community. Representing a person's social interactions in terms of their cores provides a powerful simplification of social systems. Viewing cores as social states lets us to think about behavior in terms of information theory and allows us to quantify the complexity of human social life.

5.1 Concluding remarks

This dissertation is but a small part of the growing corpus of knowledge about networks. The science of networks is still a relatively new science, one that cannot be categorized as a subfield of any traditional science, rather it is an interdisciplinary endeavor that during the past 15 years rapidly has gained the attention of the broader scientific community and society in general. Nowadays there are a dozen or so yearly conferences, workshops and schools that focus on networks. There are three entire journals devoted to the science of networks and every year multiple new books are published. During the last years we have seen an increase in the number of universities that offer network science courses and it is now even possible to earn a PhD in networks science. So it is tempting to ask: Where are we currently standing? Can we use networks to understand the complexity of our world? Can we use networks in order to design better, more stable and efficient systems? We are not there yet, not by a long shot as the financial crisis in the late 2000's clearly demonstrated. Our world is stunningly complex but also fragile. It is one thing is to understand how the collapse of *Lehman Brothers* brought down the world's financial system and how the subsequent cascade swept around the world. But it is an entirely other thing to say under which precise conditions this can happen again and when. Nevertheless we have seen considerable progress, as network science has provided us with new ways of thinking about traditional problems; ways that might lead to novel findings. Almost 300 years after *Euler* laid the foundations to graph theory, we have just begun our quest of understanding the world, and have a long journey ahead. Or as was phrased by Winston Churchill after the battle of El Alamein:

"Now this is not the end. It is not even the beginning of the end. But it is, perhaps the end of the beginning."

— Winston Churchill, November 10th, 1942

References

- Acuna, D. E., Allesina, S., and Kording, K. P. (2012). Future impact: Predicting scientific success. *Nature*, 489(7415):201–202.
- Adamic, L. A. and Huberman, B. A. (2000). Power-law distribution of the world wide web. *Science*, 287(5461):2115–2115.
- Aharony, N., Pan, W., Ip, C., Khayal, I., and Pentland, A. (2011). Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing*, 7(6):643–659.
- Ahn, Y.-Y., Bagrow, J. P., and Lehmann, S. (2010). Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764.
- Aiello, W., Chung, F., and Lu, L. (2000). A random graph model for massive graphs. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 171–180. ACM.
- Alberich, R., Miro-Julia, J., and Rossello, F. (2002). Marvel universe looks almost like a real social network. *ArXiv preprint cond-mat/0202174*.
- Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47.
- Albert, R., Jeong, H., and Barabási, A.-L. (1999). Internet: Diameter of the world-wide web. *Nature*, 401(6749):130–131.
- Albert, R., Jeong, H., and Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382.
- Amaral, L. A. N., Scala, A., Barthélemy, M., and Stanley, H. E. (2000). Classes of small-world networks. *Proceedings of the National Academy of Sciences of the United States of America*, 97(21):11149–11152.
- Anderson, P. W. et al. (1972). More is different. *Science*, 177(4047):393–396.
- Aral, S. and Walker, D. (2012). Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341.

- Backstrom, L., Boldi, P., Rosa, M., Ugander, J., and Vigna, S. (2012). Four degrees of separation. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 33–42. ACM.
- Bagrow, J. P., Wang, D., and Barabasi, A.-L. (2011). Collective response of human populations to large-scale emergencies. *PLoS ONE*, 6(3):e17680.
- Baird, D. and Ulanowicz, R. E. (1989). The seasonal dynamics of the chesapeake bay ecosystem. *Ecological Monographs*, pages 329–364.
- Bajardi, P., Barrat, A., Natale, F., Savini, L., and Colizza, V. (2011). Dynamical patterns of cattle trade movements. *PLoS ONE*, 6(5):e19869.
- Barabasi, A.-L. (2002). *Linked: How everything is connected to everything else and what it means*. Plume Editors.
- Barabasi, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211.
- Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.
- Barabási, A.-L., Albert, R., and Jeong, H. (2000). Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A: Statistical Mechanics and its Applications*, 281(1):69–77.
- Barrat, A., Barthelemy, M., Pastor-Satorras, R., and Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3747–3752.
- Barrat, A. and Weigt, M. (2000). On the properties of small-world network models. *The European Physical Journal B-Condensed Matter and Complex Systems*, 13(3):547–560.
- Bassett, D. S., Wymbs, N. F., Porter, M. A., Mucha, P. J., Carlson, J. M., and Grafton, S. T. (2011). Dynamic reconfiguration of human brain networks during learning. *Proceedings of the National Academy of Sciences of the United States of America*, 108(18):7641–7646.
- Bearman, P. S., Moody, J., and Stovel, K. (2004). Chains of affection: The structure of adolescent romantic and sexual networks. *American Journal of Sociology*, 110(1):44–91.
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.
- Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics reports*, 424(4):175–308.

- Bollobás, B. and Riordan, O. (2004). The diameter of a scale-free random graph. *Combinatorica*, 24(1):5–34.
- Brockmann, D. and Helbing, D. (2013). The hidden geometry of complex, network-driven contagion phenomena. *Science*, 342(6164):1337–1342.
- Buchanan, M. (2003). *Nexus: Small worlds and the groundbreaking theory of networks*. WW Norton & Company.
- Calabrese, F., Smoreda, Z., Blondel, V. D., and Ratti, C. (2011). Interplay between telecommunications and face-to-face interactions: A study using mobile phone data. *PLoS ONE*, 6(7):e20814.
- Calcagno, V., Demoinet, E., Gollner, K., Guidi, L., Ruths, D., and De Mazancourt, C. (2012). Flows of research manuscripts among scientific journals reveal hidden submission patterns. *Science*, 338(6110):1065–1069.
- Callaway, E. (2015). Young scientists go for fresh ideas. *Nature*, 518(7539):283–284.
- Candia, J., González, M. C., Wang, P., Schoenharl, T., Madey, G., and Barabási, A.-L. (2008). Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015.
- Cattuto, C., Van den Broeck, W., Barrat, A., Colizza, V., Pinton, J.-F., and Vespignani, A. (2010). Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLoS ONE*, 5(7):e11596.
- Centola, D. (2010). The spread of behavior in an online social network experiment. *Science*, 329(5996):1194–1197.
- Chen, J. and Yuan, B. (2006). Detecting functional modules in the yeast protein–protein interaction network. *Bioinformatics*, 22(18):2283–2290.
- Cho, E., Myers, S. A., and Leskovec, J. (2011). Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1082–1090. ACM.
- Clauset, A. and Eagle, N. (2012). Persistence and periodicity in a dynamic proximity network. *ArXiv preprint arXiv:1211.7343*.
- Clauset, A., Newman, M. E., and Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, 70(6):066111.
- Cohen, J. E., Briand, F., and Newman, C. M. (1990). *Community food webs: Data and Theory*. Springer-Verlag Berlin Heidelberg.
- Cohen, R. and Havlin, S. (2003). Scale-free networks are ultrasmall. *Physical Review Letters*, 90(5):058701.

- Cooley, C. H. (1910). *Social organization: A study of the larger mind*. Charles Scribner's Sons, New York.
- Costa, L. d. F., Rodrigues, F. A., Travieso, G., and Villas Boas, P. R. (2007). Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242.
- Crandall, D. J., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D., and Kleinberg, J. (2010). Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences of the United States of America*, 107(52):22436–22441.
- Cuttone, A., Lehmann, S., and Larsen, J. E. (2014). Inferring human mobility from sparse low accuracy mobile sensing data. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 995–1004. ACM.
- Daqing, L., Kosmidis, K., Bunde, A., and Havlin, S. (2011). Dimension of spatially embedded networks. *Nature Physics*, 7(6):481–484.
- Davis, A., Gardner, B. B., and Gardner, M. R. (1941). *Deep south*. University of Chicago Press.
- Davis, G. F., Yoo, M., and Baker, W. E. (2003). The small world of the american corporate elite, 1982–2001. *Strategic Organization*, 1(3):301–326.
- De Domenico, M., Lima, A., and Musolesi, M. (2013a). Interdependence and predictability of human mobility and social interactions. *Pervasive and Mobile Computing*, 9(6):798–807.
- De Domenico, M., Solé-Ribalta, A., Cozzo, E., Kivelä, M., Moreno, Y., Porter, M. A., Gómez, S., and Arenas, A. (2013b). Mathematical formulation of multilayer networks. *Physical Review X*, 3(4):041022.
- de Lichtenberg, U., Jensen, L. J., Brunak, S., and Bork, P. (2005). Dynamic complex formation during the yeast cell cycle. *Science*, 307(5710):724–727.
- de Montjoye, Y.-A., Stopczynski, A., Shmueli, E., Pentland, A., and Lehmann, S. (2014). The strength of the strongest ties in collaborative problem solving. *Scientific Reports*, 4.
- de Ruiter, P. C., Wolters, V., and Moore, J. C. (2005). *Dynamic food webs: Multispecies assemblages, ecosystem development and environmental change*, volume 3. Academic Press.
- de Solla Price, D. (1963). *Little science, big science*. Columbia University, New York.
- de Solla Price, D. (1965). Networks of scientific papers. *Science*, 149(3683):510–515.

- de Solla Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306.
- Derényi, I., Palla, G., and Vicsek, T. (2005). Clique percolation in random networks. *Physical Review Letters*, 94(16):160202.
- Dimitriadis, S. I., Laskaris, N. A., Tsirka, V., Vourkas, M., Micheloyannis, S., and Fotopoulos, S. (2010). Tracking brain dynamics via time-dependent network analysis. *Journal of Neuroscience Methods*, 193(1):145–155.
- Dodds, P. S., Muhamad, R., and Watts, D. J. (2003). An experimental study of search in global social networks. *Science*, 301(5634):827–829.
- Dodds, P. S. and Rothman, D. H. (1999). Unified view of scaling laws for river networks. *Physical Review E*, 59(5):4865.
- Dorogovtsev, S. N. and Mendes, J. F. (2002). Evolution of networks. *Advances in Physics*, 51(4):1079–1187.
- Dourisboure, Y., Geraci, F., and Pellegrini, M. (2007). Extraction and classification of dense communities in the web. In *Proceedings of the 16th International Conference on World Wide Web*, pages 461–470. ACM.
- Dunne, J. A., Williams, R. J., and Martinez, N. D. (2002). Food-web structure and network theory: the role of connectance and size. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12917–12922.
- Eagle, N. and Pentland, A. (2006). Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268.
- Eagle, N. and Pentland, A. S. (2009). Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066.
- Eagle, N., Pentland, A. S., and Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America*, 106(36):15274–15278.
- Easley, D. and Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press.
- Ebel, H., Mielsch, L.-I., and Bornholdt, S. (2002). Scale-free topology of e-mail networks. *Physical Review E*, 66(3):035103.
- Eckmann, J.-P., Moses, E., and Sergi, D. (2004). Entropy of dialogues creates coherent structures in e-mail traffic. *Proceedings of the National Academy of Sciences of the United States of America*, 101(40):14333–14337.

- Elowitz, M. B. and Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–338.
- Erdős, P. and Rényi, A. (1959). On random graphs, I. *Publ. Math. Debrecen*, 6:290–297.
- Faloutsos, M., Faloutsos, P., and Faloutsos, C. (1999). On power-law relationships of the internet topology. In *ACM SIGCOMM Computer Communication Review*, volume 29, pages 251–262. ACM.
- Fano, R. M. (1961). *Transmission of information: A statistical theory of communications*. MIT Press.
- Flake, G. W., Lawrence, S., Giles, C. L., and Coetzee, F. M. (2002). Self-organization and identification of web communities. *Computer*, 35(3):66–70.
- Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3):75–174.
- Fortunato, S. and Barthélemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences of the United States of America*, 104(1):36–41.
- Gao, Y., Kontoyiannis, I., and Bienenstock, E. (2008). Estimating the entropy of binary time series: Methodology, some theory and a simulation study. *Entropy*, 10(2):71–99.
- Gautreau, A., Barrat, A., and Barthélemy, M. (2009). Microdynamics in stationary complex networks. *Proceedings of the National Academy of Sciences of the United States of America*, 106(22):8847–8852.
- Gauvin, L., Panisson, A., and Cattuto, C. (2014). Detecting the community structure and activity patterns of temporal networks: a non-negative tensor factorization approach. *PLoS ONE*, 9(1):e86028.
- Giancarlo, R., Scaturro, D., and Utro, F. (2008). Computational cluster validation for microarray data analysis: experimental assessment of cleft, consensus clustering, figure of merit, gap statistics and model explorer. *BMC Bioinformatics*, 9(1):462.
- Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–7826.
- Goffman, E. (1967). *Interaction Ritual: Essays in Face to Face Behavior*. AldineTransaction.
- Goh, K.-I. and Barabási, A.-L. (2008). Burstiness and memory in complex systems. *EPL (Europhysics Letters)*, 81(4):48002.
- Goldenfeld, N. and Kadanoff, L. P. (1999). Simple lessons from complexity. *Science*, 284(5411):87–89.

- Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.
- Good, B. H., de Montjoye, Y.-A., and Clauset, A. (2010). Performance of modularity maximization in practical contexts. *Physical Review E*, 81(4):046106.
- Gower, J. C. and Ross, G. (1969). Minimum spanning trees and single linkage cluster analysis. *Applied Statistics*, pages 54–64.
- Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, pages 1360–1380.
- Guare, J. (1990). *Six degrees of separation: A play*. Vintage Books, New York.
- Guelzim, N., Bottani, S., Bourguin, P., and Képès, F. (2002). Topological and causal structure of the yeast transcriptional regulatory network. *Nature Genetics*, 31(1):60–63.
- Guimera, R., Mossa, S., Turtschi, A., and Amaral, L. N. (2005). The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(22):7794–7799.
- Hall, E. T. (1969). *The hidden dimension*, volume 1990. Anchor Books New York.
- Holme, P. (2003). Network dynamics of ongoing social relationships. *EPL (Europhysics Letters)*, 64(3):427.
- Holme, P. (2005). Network reachability of real-world contact sequences. *Physical Review E*, 71(4):046119.
- Holme, P., Kim, B. J., Yoon, C. N., and Han, S. K. (2002). Attack vulnerability of complex networks. *Physical Review E*, 65(5):056109.
- Holme, P. and Saramäki, J. (2012). Temporal networks. *Physics Reports*, 519(3):97–125.
- Homans, G. C. (1951). *The human group*. Routledge.
- Iribarren, J. L. and Moro, E. (2009). Impact of human activity patterns on the dynamics of information diffusion. *Physical Review Letters*, 103(3):038702.
- Isella, L., Romano, M., Barrat, A., Cattuto, C., Colizza, V., Van den Broeck, W., Gesualdo, F., Pandolfi, E., Ravà, L., Rizzo, C., et al. (2011). Close encounters in a pediatric ward: measuring face-to-face proximity and mixing patterns with wearable sensors. *PLoS ONE*, 6(2):e17144.
- Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, 98(8):4569–4574.

- ITU (2014). Key ICT indicators for developed and developing countries and the world (totals and penetration rates). Technical report, International Telecommunications Union.
- Jaffe, A. B. and Trajtenberg, M. (2002). *Patents, citations, and innovations: A window on the knowledge economy*. MIT press.
- Jensen, B. S., Larsen, J. E., Jensen, K., Larsen, J., and Hansen, L. K. (2010). Estimating human predictability from mobile sensor data. In *Machine Learning for Signal Processing (MLSP), 2010 IEEE International Workshop on*, pages 196–201. IEEE.
- Jeong, H., Mason, S. P., Barabási, A.-L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833):41–42.
- Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabási, A.-L. (2000). The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654.
- Jo, H.-H., Pan, R. K., and Kaski, K. (2011). Emergence of bursts and communities in evolving weighted networks. *PLoS ONE*, 6(8):e22687.
- Jonsson, P. E., Cavanna, T., Zicha, D., and Bates, P. A. (2006). Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. *BMC Bioinformatics*, 7(1):2.
- Jordano, P., Bascompte, J., and Olesen, J. M. (2003). Invariant properties in coevolutionary networks of plant–animal interactions. *Ecology Letters*, 6(1):69–81.
- Kalmijn, M. (1998). Inter-marriage and homogamy: Causes, patterns, trends. *Annual Review of Sociology*, pages 395–421.
- Karsai, M., Kivela, M., Pan, R. K., Kaski, K., Kertész, J., Barabási, A.-L., and Saramäki, J. (2011). Small but slow world: How network topology and burstiness slow down spreading. *Physical Review E*, 83(2):025102.
- Karsai, M., Perra, N., and Vespignani, A. (2014). Time varying networks and the weakness of strong ties. *Scientific Reports*, 4.
- Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43.
- Kempe, D., Kleinberg, J., and Kumar, A. (2000). Connectivity and inference problems for temporal networks. In *Proceedings of the Thirty-second Annual ACM Symposium on Theory of Computing*, pages 504–513. ACM.
- Kernighan, B. W. and Lin, S. (1970). An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*, 49(2):291–307.
- Kiukkonen, N., Blom, J., Dousse, O., Gatica-Perez, D., and Laurila, J. (2010). Towards rich mobile phone datasets: Lausanne data collection campaign. *Proc. ICPS, Berlin*.

- Kleinberg, J. M. (2000). Navigation in a small world. *Nature*, 406(6798):845–845.
- Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. S. (1999). The web as a graph: measurements, models, and methods. In *Computing and Combinatorics*, pages 1–17. Springer.
- Klug, M. and Bagrow, J. P. (2014). Understanding the group dynamics and success of teams. *ArXiv preprint arXiv:1407.2893*.
- Kontoyiannis, I., Algoet, P. H., Suhov, Y. M., and Wyner, A. J. (1998). Nonparametric entropy estimation for stationary processes and random fields, with applications to english text. *Information Theory, IEEE Transactions on*, 44(3):1319–1327.
- Kossinets, G. (2006). Effects of missing data in social networks. *Social Networks*, 28(3):247–268.
- Kossinets, G. and Watts, D. J. (2006). Empirical analysis of an evolving social network. *Science*, 311(5757):88–90.
- Krapivsky, P. L. and Redner, S. (2001). Organization of growing random networks. *Physical Review E*, 63(6):066123.
- Krapivsky, P. L. and Redner, S. (2002). A statistical physics perspective on web growth. *Computer Networks*, 39(3):261–276.
- Krause, A. E., Frank, K. A., Mason, D. M., Ulanowicz, R. E., and Taylor, W. W. (2003). Compartments revealed in food-web structure. *Nature*, 426(6964):282–285.
- Krebs, V. E. (2002). Mapping networks of terrorist cells. *Connections*, 24(3):43–52.
- Krings, G., Karsai, M., Bernhardsson, S., Blondel, V. D., and Saramäki, J. (2012). Effects of time window size and placement on the structure of an aggregated communication network. *EPJ Data Science*, 1(4):1–16.
- Laksanasopin, T., Guo, T. W., Nayak, S., Sridhara, A. A., Xie, S., Olowookere, O. O., Cadinu, P., Meng, F., Chee, N. H., Kim, J., et al. (2015). A smartphone dongle for diagnosis of infectious diseases at the point of care. *Science Translational Medicine*, 7(273):273re1–273re1.
- Lamport, L. (1978). Time, clocks, and the ordering of events in a distributed system. *Communications of the ACM*, 21(7):558–565.
- Lancichinetti, A., Radicchi, F., Ramasco, J. J., and Fortunato, S. (2011). Finding statistically significant communities in networks. *PLoS ONE*, 6(4):e18961.
- Lane, N. D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., and Campbell, A. T. (2010). A survey of mobile phone sensing. *Communications Magazine, IEEE*, 48(9):140–150.

- Latora, V. and Marchiori, M. (2001). Efficient behavior of small-world networks. *Physical Review Letters*, 87(19):198701.
- Latora, V. and Marchiori, M. (2002). Is the boston subway a small-world network? *Physica A: Statistical Mechanics and its Applications*, 314(1):109–113.
- Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Van Alstyne, M. (2009). Computational social science. *Science*, 323(5915):721–723.
- Lehmann, S., Jackson, A. D., and Lautrup, B. E. (2006). Measures for measures. *Nature*, 444(7122):1003–1004.
- Leskovec, J. and Horvitz, E. (2008). Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th International Conference on World Wide Web*, pages 915–924. ACM.
- Lewis, J. (2003). Autoinhibition with transcriptional delay: a simple mechanism for the zebrafish somitogenesis oscillator. *Current Biology*, 13(16):1398–1408.
- Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., and Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using facebook. com. *Social Networks*, 30(4):330–342.
- Liljeros, F., Edling, C. R., Amaral, L. A. N., Stanley, H. E., and Åberg, Y. (2001). The web of human sexual contacts. *Nature*, 411(6840):907–908.
- Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of Washington Academy Sciences*.
- Lu, X., Bengtsson, L., and Holme, P. (2012). Predictability of population displacement after the 2010 haiti earthquake. *Proceedings of the National Academy of Sciences of the United States of America*, 109(29):11576–11581.
- Łuczak, T. (1989). Sparse random graphs with a given degree sequence. In *Proceedings of the Symposium on Random Graphs, Poznan*, pages 165–182.
- Lusseau, D., Schneider, K., Boisseau, O. J., Haase, P., Slooten, E., and Dawson, S. M. (2003). The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, 54(4):396–405.
- Malmgren, R. D., Ottino, J. M., and Amaral, L. A. N. (2010). The role of mentorship in protégé performance. *Nature*, 465(7298):622–626.
- Malmgren, R. D., Stouffer, D. B., Campanharo, A. S., and Amaral, L. A. N. (2009). On universality in human correspondence activity. *Science*, 325(5948):1696–1700.
- Mariolis, P. (1975). Interlocking directorates and control of corporations: The theory of bank control. *Social Science Quarterly*, pages 425–439.

- Maritan, A., Rinaldo, A., Rigon, R., Giacometti, A., and Rodríguez-Iturbe, I. (1996). Scaling laws for river networks. *Physical Review E*, 53(2):1510.
- Marsden, P. V. (1990). Network data and measurement. *Annual Review of Sociology*, pages 435–463.
- Maslov, S. and Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913.
- McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, pages 415–444.
- Milgram, S. (1967). The small world problem. *Psychology Today*, 2(1):60–67.
- Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science*, 7(3):221–237.
- Milojević, S. (2014). Principles of scientific research team formation and evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 111(11):3984–3989.
- Miritello, G., Lara, R., Cebrian, M., and Moro, E. (2013). Limited communication capacity unveils strategies for human interaction. *Scientific Reports*, 3.
- Miritello, G., Moro, E., and Lara, R. (2011). Dynamical strength of social ties in information spreading. *Physical Review E*, 83(4):045102.
- Molloy, M. and Reed, B. (1995). A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms*, 6(2-3):161–180.
- Moreno, J. L. and Jennings, H. H. (1934). *Who shall survive?: A new approach to the problem of human interrelations*. Beacon House, Beacon, NY.
- Mucha, P. J., Richardson, T., Macon, K., Porter, M. A., and Onnela, J.-P. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–878.
- Nelson, D., Ihekwebaba, A., Elliott, M., Johnson, J., Gibney, C., Foreman, B., Nelson, G., See, V., Horton, C., Spiller, D., et al. (2004). Oscillations in NF- κ B signaling control the dynamics of gene expression. *Science*, 306(5696):704–708.
- Newman, M., Barabási, A.-L., and Watts, D. J. (2006). *The structure and dynamics of networks*. Princeton University Press.
- Newman, M. E. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, 98(2):404–409.
- Newman, M. E. (2002). Assortative mixing in networks. *Physical Review Letters*, 89(20):208701.

- Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.
- Newman, M. E. (2004a). Analysis of weighted networks. *Physical Review E*, 70(5):056131.
- Newman, M. E. (2004b). Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(6):066133.
- Newman, M. E. (2005). Power laws, pareto distributions and zipf’s law. *Contemporary Physics*, 46(5):323–351.
- Newman, M. E. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):036104.
- Newman, M. E. (2008). The mathematics of networks. *The New Palgrave Encyclopedia of Economics*, 2:1–12.
- Newman, M. E., Forrest, S., and Balthrop, J. (2002). Email networks and the spread of computer viruses. *Physical Review E*, 66(3):035101.
- Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113.
- Newman, M. E. and Park, J. (2003). Why social networks are different from other types of networks. *Physical Review E*, 68(3):036122.
- Oliveira, J. G. and Barabási, A.-L. (2005). Human dynamics: Darwin and einstein correspondence patterns. *Nature*, 437(7063):1251–1251.
- Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., and Barabási, A.-L. (2007). Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences of the United States of America*, 104(18):7332–7336.
- Padgett, J. F. and Ansell, C. K. (1993). Robust action and the rise of the medici, 1400-1434. *American Journal of Sociology*, pages 1259–1319.
- Pahl-Wostl, C. (1995). *The dynamic nature of ecosystems: Chaos and order entwined*. Wiley Chichester.
- Palla, G., Barabási, A.-L., and Vicsek, T. (2007). Quantifying social group evolution. *Nature*, 446(7136):664–667.
- Palla, G., Derényi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818.
- Pan, R. K. and Saramäki, J. (2011). Path lengths, correlations, and centrality in temporal networks. *Physical Review E*, 84(1):016105.

- Pareto, V. (1896). *Cours d'économie politique: professé à l'Université de Lausanne*, volume 1. F. Rouge.
- Pentland, A. (2012). The new science of building great teams. *Harvard Business Review*, 90(4):60–69.
- Pentland, A. S. (2008). *Honest Signals: How They Shape Our World*. The MIT press.
- Petri, G., Expert, P., Turkheimer, E., Carhart-Harris, R., Nutt, D., Hellyer, P., and Vaccarino, F. (2014). Homological scaffolds of brain functional networks. *Journal of The Royal Society Interface*, 11(101):20140873.
- Pimm, S. L. (1979). The structure of food webs. *Theoretical Population Biology*, 16(2):144–158.
- Pons, P. and Latapy, M. (2011). Post-processing hierarchical community structures: Quality improvements and multi-scale view. *Theoretical Computer Science*, 412(8):892–900.
- Porta, S., Crucitti, P., and Latora, V. (2006). The network analysis of urban streets: a dual approach. *Physica A: Statistical Mechanics and its Applications*, 369(2):853–866.
- Rapoport, A. and Horvath, W. J. (1961). A study of a large sociogram. *Behavioral Science*, 6(4):279–291.
- Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555.
- Rice, S. A. (1927). The identification of blocs in small political bodies. *American Political Science Review*, 21(03):619–627.
- Riolo, C. S., Koopman, J. S., and Chick, S. E. (2001). Methods and measures for the description of epidemiologic contact networks. *Journal of Urban Health*, 78(3):446–457.
- Rives, A. W. and Galitski, T. (2003). Modular organization of cellular networks. *Proceedings of the National Academy of Sciences of the United States of America*, 100(3):1128–1133.
- Rosvall, M. and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, 105(4):1118–1123.
- Rosvall, M., Trusina, A., Minnhagen, P., and Sneppen, K. (2005). Networks and cities: An information perspective. *Physical Review Letters*, 94(2):028701.

- Rothlisberger, F. J. and Dickson, W. J. (1939). *Management and the worker: An account of a research program conducted by the Western Electric Company, Hawthorne Works*. Harvard University Press.
- Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, 31(4):581–603.
- Salathé, M., Kazandjieva, M., Lee, J. W., Levis, P., Feldman, M. W., and Jones, J. H. (2010). A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences of the United States of America*, 107(51):22020–22025.
- Sarigöl, E., Pfitzner, R., Scholtes, I., Garas, A., and Schweitzer, F. (2014). Predicting scientific success based on coauthorship networks. *EPJ Data Science*, 3(1):1–16.
- Schich, M., Song, C., Ahn, Y.-Y., Mirsky, A., Martino, M., Barabási, A.-L., and Helbing, D. (2014). A network framework of cultural history. *Science*, 345(6196):558–562.
- Schweitzer, F., Fagiolo, G., Sornette, D., Vega-Redondo, F., Vespignani, A., and White, D. R. (2009). Economic networks: The new challenges. *Science*, 325:422–425.
- Scott, J. and Carrington, P. J. (2011). *The SAGE handbook of social network analysis*. SAGE publications.
- Sen, P., Dasgupta, S., Chatterjee, A., Sreeram, P., Mukherjee, G., and Manna, S. (2003). Small-world properties of the indian railway network. *Physical Review E*, 67(3):036106.
- Shalizi, C. R. and Thomas, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, 40(2):211–239.
- Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.
- Shen, H.-W. and Barabási, A.-L. (2014). Collective credit allocation in science. *Proceedings of the National Academy of Sciences of the United States of America*, 111(34):12325–12330.
- Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of escherichia coli. *Nature genetics*, 31(1):64–68.
- Sigman, M. and Cecchi, G. A. (2002). Global organization of the wordnet lexicon. *Proceedings of the National Academy of Sciences of the United States of America*, 99(3):1742–1747.
- Simmel, G. (1950). Quantitative aspects of the group. *The Sociology of Georg Simmel*, pages 87–177.

- Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, pages 425–440.
- Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. (2010). Limits of predictability in human mobility. *Science*, 327(5968):1018–1021.
- Spirin, V. and Mirny, L. A. (2003). Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences of the United States of America*, 100(21):12123–12128.
- Sporns, O. (2002). Network analysis, complexity, and brain function. *Complexity*, 8(1):56–60.
- Stehlé, J., Voirin, N., Barrat, A., Cattuto, C., Colizza, V., Isella, L., Régis, C., Pinton, J.-F., Khanafer, N., Van den Broeck, W., et al. (2011a). Simulation of an SEIR infectious disease model on the dynamic contact network of conference attendees. *BMC Medicine*, 9(1):87.
- Stehlé, J., Voirin, N., Barrat, A., Cattuto, C., Isella, L., Pinton, J.-F., Quaggiotto, M., Van den Broeck, W., Régis, C., Lina, B., et al. (2011b). High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS ONE*, 6(8):e23176.
- Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S., and Gilles, E. D. (2002). Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 420(6912):190–193.
- Stopczynski, A., Sekara, V., Sapiezynski, P., Cuttone, A., Madsen, M. M., Larsen, J. E., and Lehmann, S. (2014a). Measuring large-scale social networks with high resolution. *PLoS ONE*, 9(4):e95978.
- Stopczynski, A., Stahlhut, C., Larsen, J. E., Petersen, M. K., and Hansen, L. K. (2014b). The smartphone brain scanner: a portable real-time neuroimaging system. *PLoS ONE*, 9(2):e86733.
- Sulo, R., Berger-Wolf, T., and Grossman, R. (2010). Meaningful selection of temporal resolution for dynamic networks. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, pages 127–136. ACM.
- Szell, M. and Thurner, S. (2013). How women organize social networks different from men. *Scientific Reports*, 3.
- Tang, J., Musolesi, M., Mascolo, C., Latora, V., and Nicosia, V. (2010). Analysing information flows and key mediators through temporal centrality metrics. In *Proceedings of the 3rd Workshop on Social Network Systems*, page 3. ACM.
- Tantipathananandh, C., Berger-Wolf, T., and Kempe, D. (2007). A framework for community identification in dynamic social networks. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 717–726. ACM.

- Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.
- Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., et al. (2000). A comprehensive analysis of protein–protein interactions in *saccharomyces cerevisiae*. *Nature*, 403(6770):623–627.
- Valencia, M., Martinerie, J., Dupont, S., and Chavez, M. (2008). Dynamic small-world behavior in functional brain networks unveiled by an event-related networks approach. *Physical Review E*, 77(5):050905.
- Vernon, M. C. and Keeling, M. J. (2009). Representing the uk's cattle herd as static and dynamic networks. *Proceedings of the Royal Society B: Biological Sciences*, 276(1656):469–476.
- Vragović, I., Louis, E., and Diaz-Guilera, A. (2005). Efficiency of informational transfer in regular and complex networks. *Physical Review E*, 71(3):036122.
- Wagner, A. and Fell, D. A. (2001). The small world inside large metabolic networks. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1478):1803–1810.
- Wang, D., Pedreschi, D., Song, C., Giannotti, F., and Barabasi, A.-L. (2011). 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.
- Wang, D., Song, C., and Barabási, A.-L. (2013). Quantifying long-term scientific impact. *Science*, 342(6154):127–132.
- Wang, P., González, M. C., Hidalgo, C. A., and Barabási, A.-L. (2009). Understanding the spreading patterns of mobile phone viruses. *Science*, 324(5930):1071–1076.
- Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.
- Wasserman, S. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge University Press.
- Watts, D. J. (1999). *Small worlds: the dynamics of networks between order and randomness*. Princeton university press.
- Watts, D. J. (2004). *Six degrees: The science of a connected age*. WW Norton & Company.
- Watts, D. J. (2007). A twenty-first century science. *Nature*, 445(7127):489–489.

- Watts, D. J., Dodds, P. S., and Newman, M. E. (2002). Identity and search in social networks. *Science*, 296(5571):1302–1305.
- Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of “small-world” networks. *Nature*, 393(6684):440–442.
- White, J., Southgate, E., Thomson, J., and Brenner, S. (1986). The structure of the nervous system of the nematode *Caenorhabditis elegans*: the mind of a worm. *Phil. Trans. R. Soc. Lond.*, 314:1–340.
- Williams, M. J. (2013). *Periodic patterns in human mobility*. PhD thesis, Cardiff University.
- Williams, R. J., Berlow, E. L., Dunne, J. A., Barabási, A.-L., and Martinez, N. D. (2002). Two degrees of separation in complex food webs. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12913–12916.
- Williams, R. J. and Martinez, N. D. (2000). Simple rules yield complex food webs. *Nature*, 404(6774):180–183.
- Wrzus, C., Hänel, M., Wagner, J., and Neyer, F. J. (2013). Social network changes and life events across the life span: A meta-analysis. *Psychological Bulletin*, 139(1):53.
- Wu, Y., Zhou, C., Xiao, J., Kurths, J., and Schellnhuber, H. J. (2010). Evidence for a bimodal distribution in human communication. *Proceedings of the National Academy of Sciences of the United States of America*, 107(44):18803–18808.
- Wuchty, S. (2009). What is a social tie? *Proceedings of the National Academy of Sciences of the United States of America*, 106(36):15099–15100.
- Wuchty, S., Jones, B. F., and Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039.
- Yook, S.-H., Jeong, H., and Barabási, A.-L. (2002). Modeling the internet’s large-scale topology. *Proceedings of the National Academy of Sciences of the United States of America*, 99(21):13382–13386.
- Zhao, K., Stehlé, J., Bianconi, G., and Barrat, A. (2011). Social network dynamics of face-to-face interactions. *Physical Review E*, 83(5):056109.
- Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley press.
- Ziv, J. and Lempel, A. (1978). Compression of individual sequences via variable-rate coding. *Information Theory, IEEE Transactions on*, 24(5):530–536.

P_{ublications}



Measuring Large-Scale Social Networks with High Resolution

Arkadiusz Stopczynski^{1*}, Vedran Sekara¹, Piotr Sapiezynski¹, Andrea Cuttone¹, Mette My Madsen³, Jakob Eg Larsen¹, Sune Lehmann^{1,2}

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

Abstract

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

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

Editor: Yamir Moreno, University of Zaragoza, Spain

Received: February 15, 2014; **Accepted:** April 2, 2014; **Published:** April 25, 2014

Copyright: © 2014 Stopczynski et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The SensibleDTU project was made possible by a Young Investigator Grant from the Villum Foundation (High Resolution Networks, awarded to SL). Scaling the project up to 1 000 individuals in 2013 was made possible by a interdisciplinary UCPH 2016 grant, Social Fabric (PI David Dreyer Lassen, SL is co-PI) focusing mainly on the social and basic science elements of the project. This grant has funded purchase of the smartphones, as well as technical personnel. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: arks@dtu.dk

Introduction

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

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

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

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

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

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

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

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

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

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

Related Work

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

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

Data collection

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

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

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

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

Data analysis

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

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

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

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

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

Health and Public Safety. Using CDRs from the period of the 2008 earthquake in Rwanda, Kapoor et al. created a model for

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

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

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

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

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

Privacy

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

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

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

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

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

Motivation

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

Multiplexity

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

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

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

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

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

Sampling

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

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

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

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

Dynamics

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

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

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

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

current methodologies, for example by adapting the link-based viewpoint to dynamical systems.

Feedback

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

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

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

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

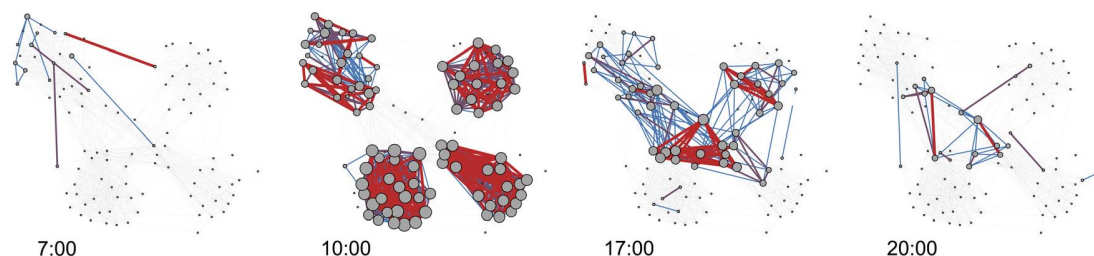


Figure 1. Dynamics of face-to-face interactions in the 2012 deployment. The participants meet in the morning, attend classes within four different study lines, and interact across majors in the evening. Edges are colored according to the frequency of observation, ranging from low (blue) to high (red).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

doi:10.1371/journal.pone.0095978.g001

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

New Science

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

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

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

Methods: Data Collection

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

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

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

Data Sources

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

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

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

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

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

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

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

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

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

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

Secondly, the inclusion of qualitative data increases the potential for interdisciplinary work between the fields of computer science and social science. Our central goal is to capture the full richness of social interactions by increasing the number of recorded communication channels. Adding a qualitative social network approach makes it possible to relate the qualitative observations to the quantitative data obtained from the mobile sensing, creating an interdisciplinary space for methods and theory. We are particularly interested in the relationship between

the observations made by the embedded anthropologist and the data recorded using questionnaires and mobile sensing, to answer questions about the elements difficult to capture using our high-resolution approach. Similarly, from the perspective of social sciences, we are able to consider what may be captured by incorporating quantitative data from mobile sensing into a qualitative data pool—and what can we learn about social networks using modern sensing technology.

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

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

Backend System

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

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

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

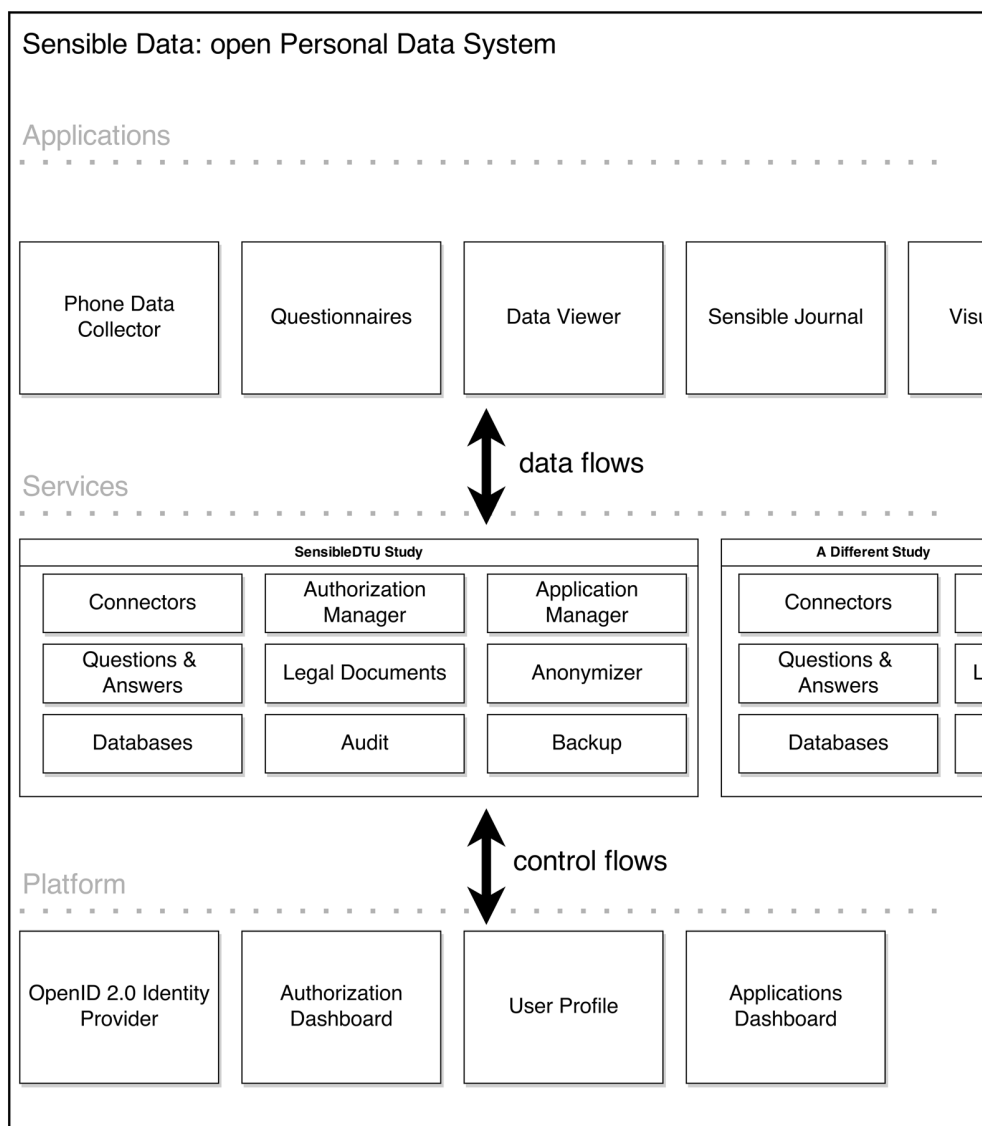


Figure 2. Sensible Data openPDS architecture. This system is used in the 2013 deployment and consists of three layers: platform, services, and applications. The platform contains element common for multiple services (in this context: studies). The studies are the deployments of particular data collection efforts. The applications are OAuth2 clients to studies and can submit and access data, based on user authorizations. doi:10.1371/journal.pone.0095978.g002

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

Deployment Methods

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

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

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

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

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

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

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

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

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

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

Methods: Privacy

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

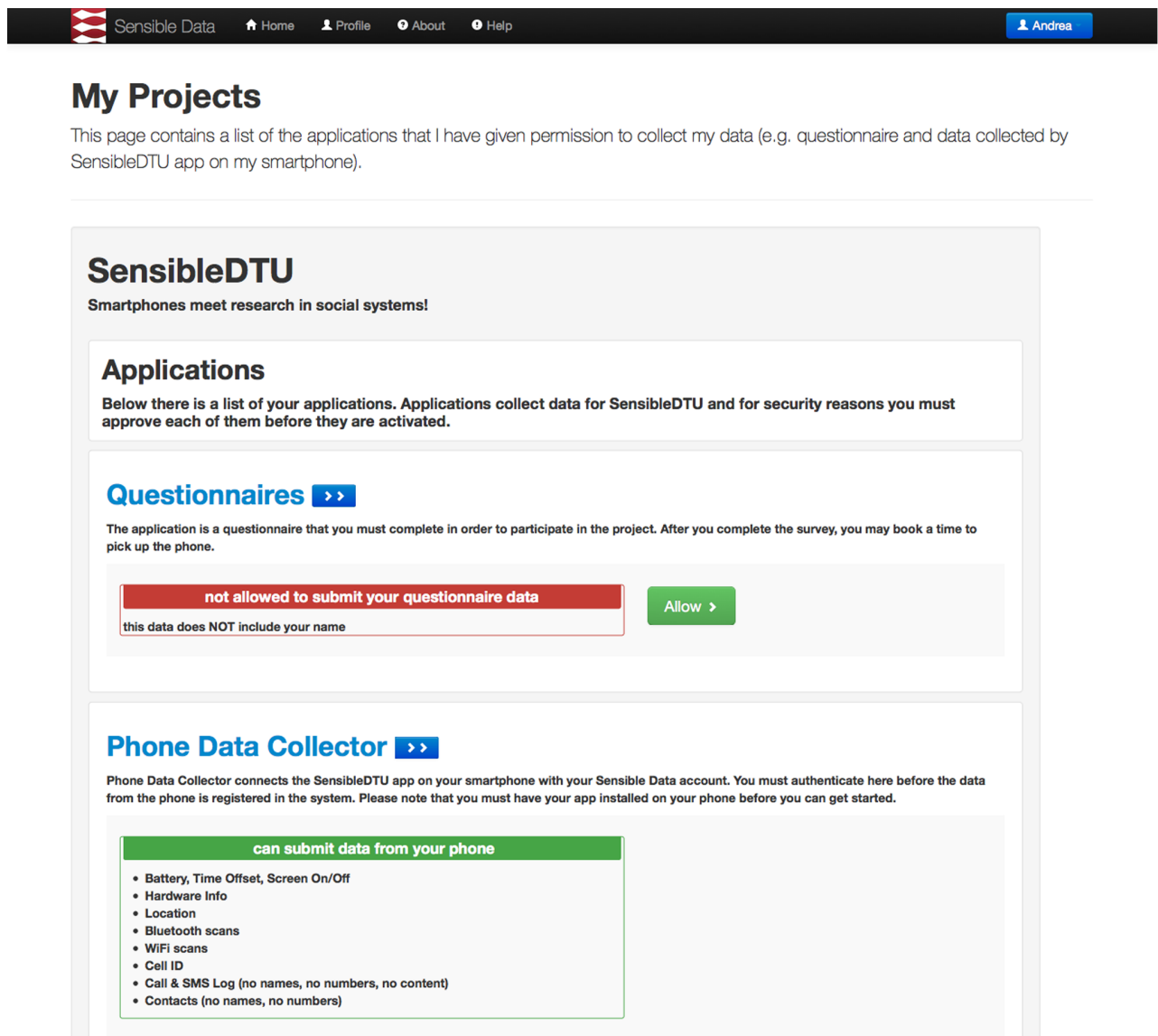
We believe that ensuring sufficient privacy for the participants, in large part, is the task of providing them with tools to align the data usage with their understanding. Such privacy tools must be of two kinds: to inform, ensuring participants understand the situation, and to control, aligning the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

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

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

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

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



Sensible Data Home Profile About Help Andrea

My Projects

This page contains a list of the applications that I have given permission to collect my data (e.g. questionnaire and data collected by SensibleDTU app on my smartphone).

SensibleDTU

Smartphones meet research in social systems!

Applications

Below there is a list of your applications. Applications collect data for SensibleDTU and for security reasons you must approve each of them before they are activated.

Questionnaires >>

The application is a questionnaire that you must complete in order to participate in the project. After you complete the survey, you may book a time to pick up the phone.

not allowed to submit your questionnaire data

this data does NOT include your name

Allow >

Phone Data Collector >>

Phone Data Collector connects the SensibleDTU app on your smartphone with your Sensible Data account. You must authenticate here before the data from the phone is registered in the system. Please note that you must have your app installed on your phone before you can get started.

can submit data from your phone

- Battery, Time Offset, Screen On/Off
- Hardware Info
- Location
- Bluetooth scans
- WiFi scans
- Cell ID
- Call & SMS Log (no names, no numbers, no content)
- Contacts (no names, no numbers)

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

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

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

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

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

Results and Discussion

As described in the previous sections, our study has collected comprehensive data about a number of aspects regarding human behavior. Below, we discuss primary data channels and report some early results and findings. The results are mainly based on the 2012 deployment due to the availability of longitudinal data.

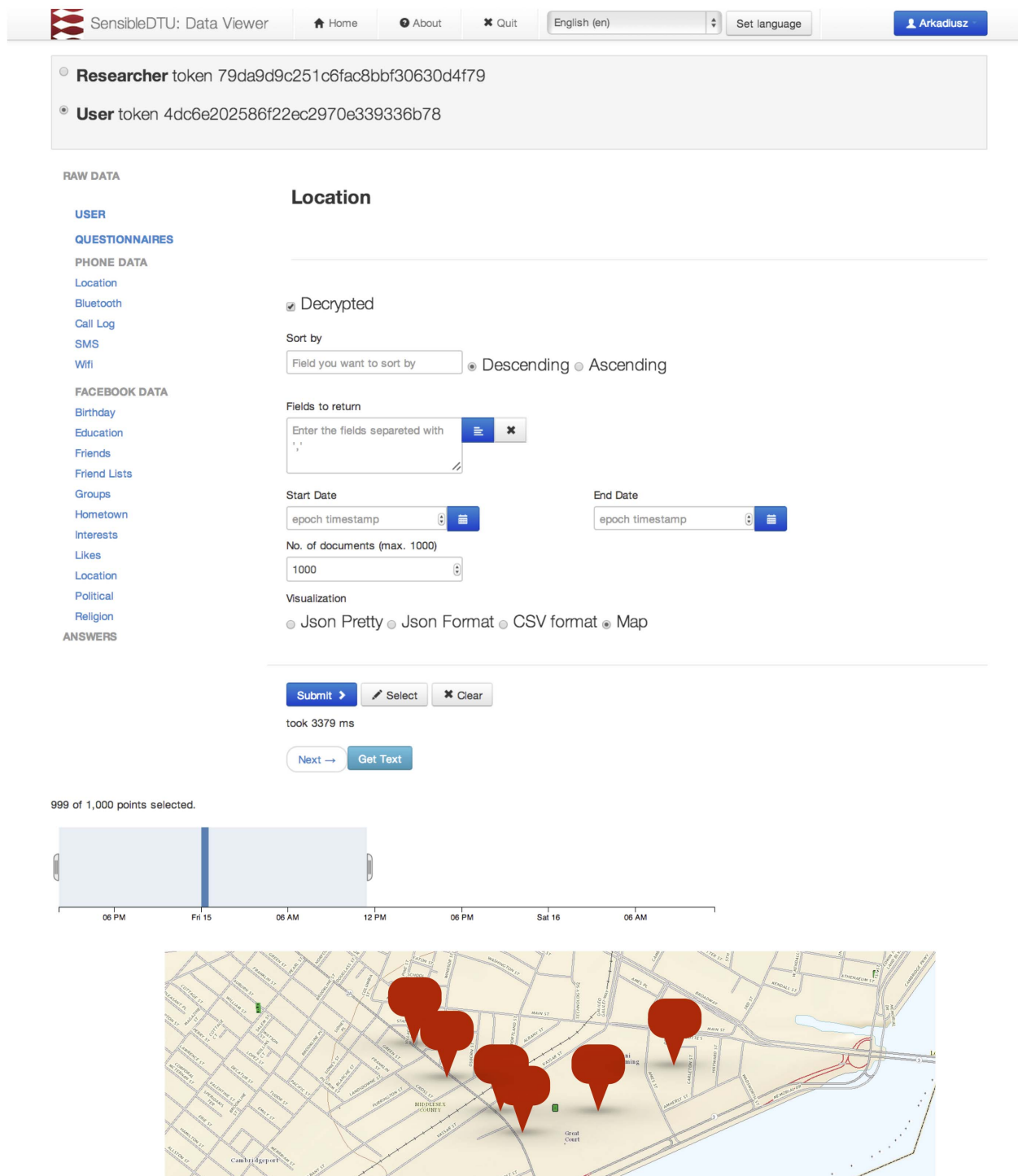


Figure 4. Data viewer application. All the collected data can be explored and accessed via an API. The API is the same for research, application, and end-user access, the endpoints are protected by OAuth2 bearer token. Map image from USGS National Map Viewer, replacing original image used in the deployed application (Google Maps).
doi:10.1371/journal.pone.0095978.g004

Bluetooth and Social Ties

Bluetooth is a wireless technology ubiquitous in modern-day mobile devices. It is used for short-range communication between devices, including smartphones, hands-free headsets, tablets, and other wearables. As the transmitters used in mobile devices are

primarily of very short range—between 5 and 10 *m* (16–33 feet)—detection of the devices of other participants (set in ‘visible’ mode) can be used as a proxy for face-to-face interactions [29]. We take the individual Bluetooth scans in the form (i, j, t, σ) , denoting that device *i* has observed device *j* at time *t* with signal strength σ .

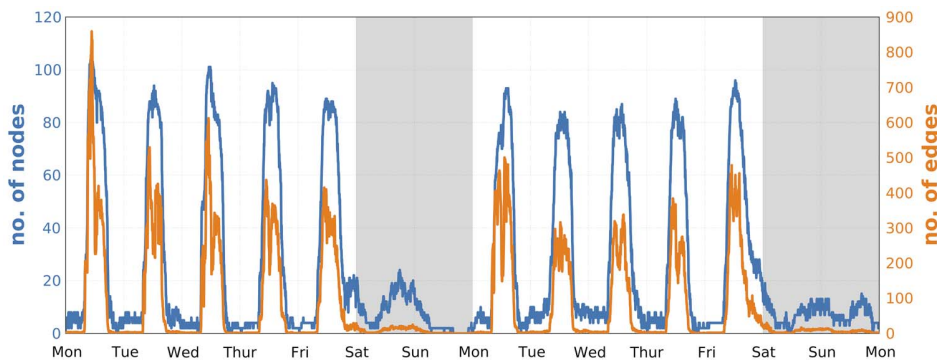


Figure 5. Weekly temporal dynamics of interactions. Face-to-face interaction patterns of participants in 5-minute time-bins over two weeks. Only active participants are included, i.e. those that have either observed another person or themselves been observed in a given time-bin. On average we observed 29 edges and 12 nodes in 5-minute time-bins and registered 10 634 unique links between participants.
doi:10.1371/journal.pone.0095978.g005

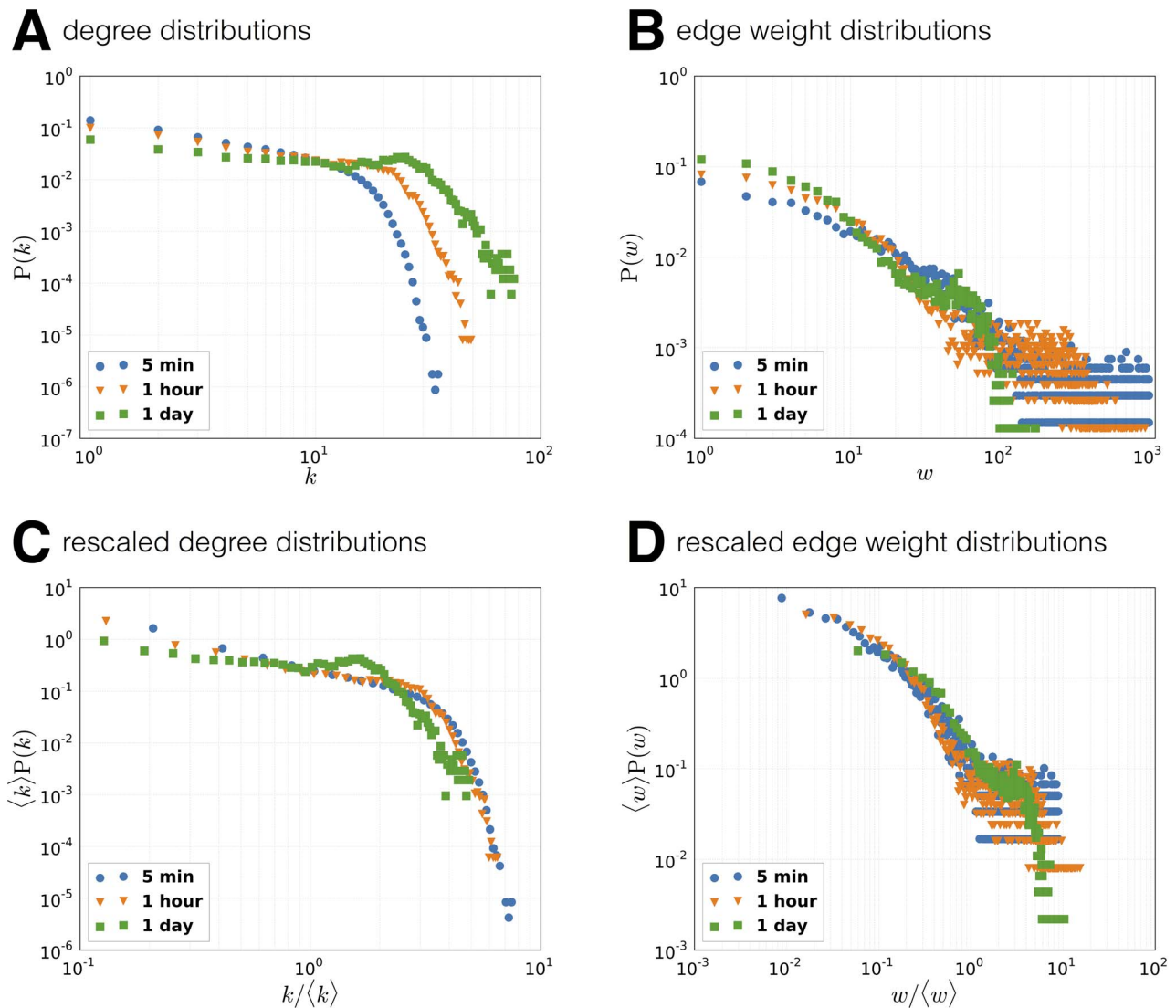


Figure 6. Face-to-face network properties at different resolution levels. Distributions are calculated by aggregating sub-distributions across temporal window. Differences in rescaled distributions suggest that social dynamics unfold on multiple timescales.
doi:10.1371/journal.pone.0095978.g006

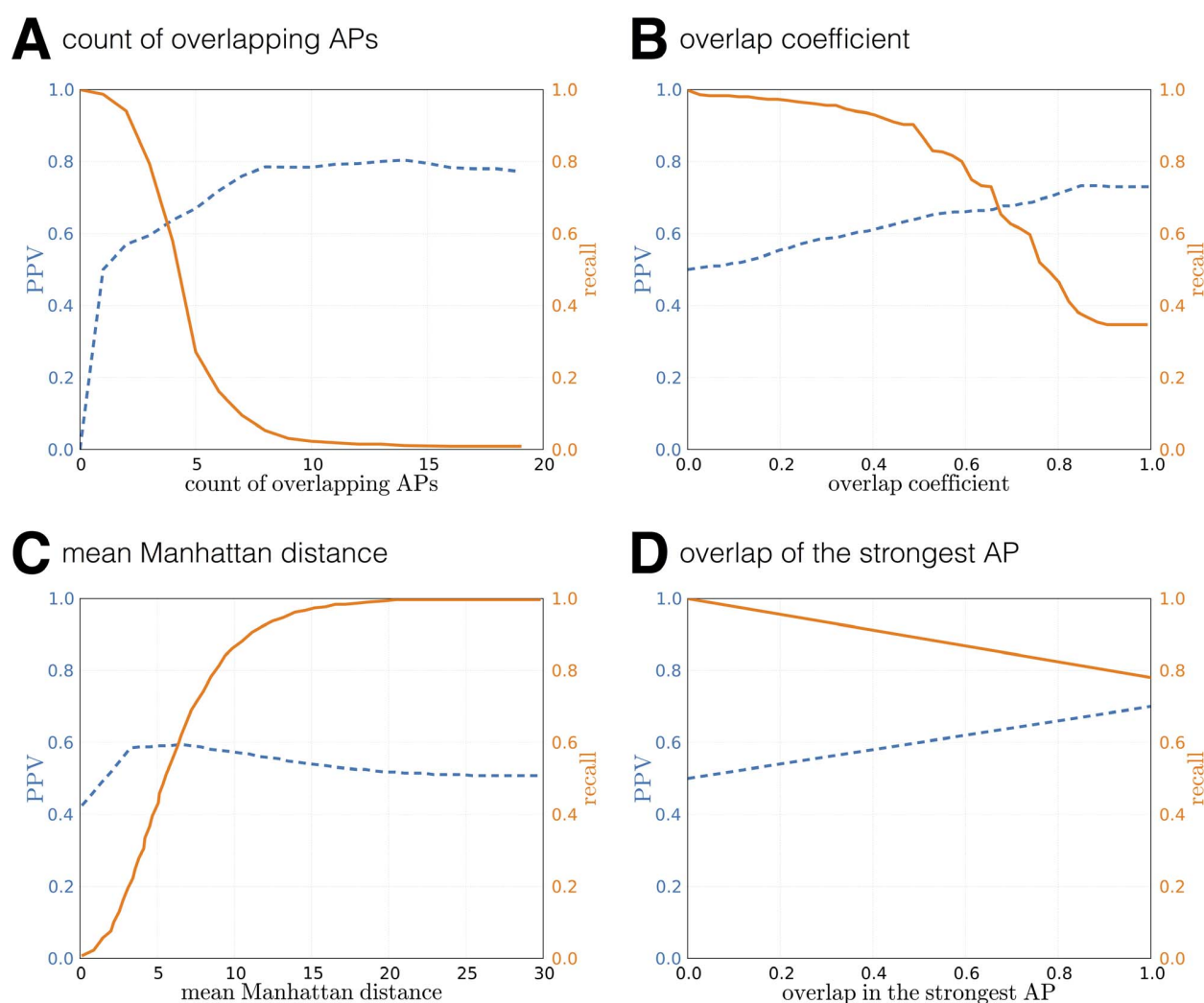


Figure 7. WiFi similarity measures. Positive predictive value (precision, ratio of number of true positives to number of positive calls, marked with dashed lines) and recall (sensitivity, fraction of retrieved positives, marked with solid lines) as functions of parameters in different similarity measures.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

Bluetooth scans do not constitute a perfect proxy for face-to-face interactions [151], since a) it is possible for people within 10 *m* radius not to interact socially, and b) it is possible to interact socially over a distance greater than 10 *m*, nevertheless, they have been successfully used for sensing social networks [31] or crowd tracking [152].

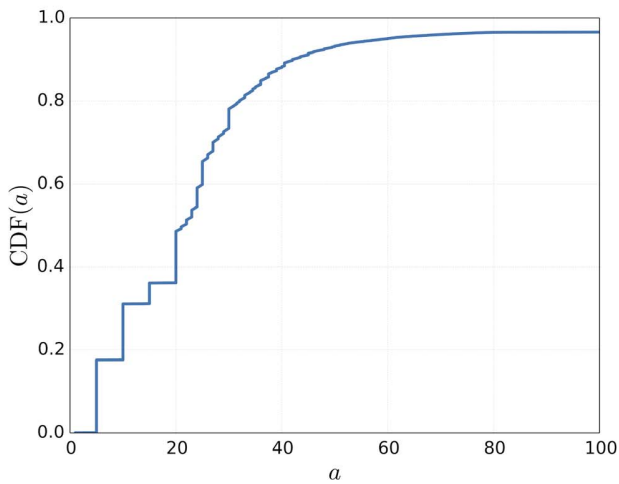
Between October 1st, 2012 and September 1st, 2013, we collected 12 623 599 Bluetooth observations in which we observed 153 208 unique devices. The scans on the participants' phones were triggered every five minutes, measured from the last time the phone was powered on. Thus, the phones scanned for Bluetooth in a desynchronized fashion, and not according to a global schedule. To account for this, when extracting interactions from the raw Bluetooth scans, we bin them into fixed-length time windows, aggregating the scans within them. The resulting adjacency matrix, $W_{\Delta t}$, does not have to be strictly symmetric, meaning that participant *i* can observe participant *j* in time-bin *t*, but not the other way around. Here we assume that Bluetooth scans do not produce false positives (devices are not discovered unless they are

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

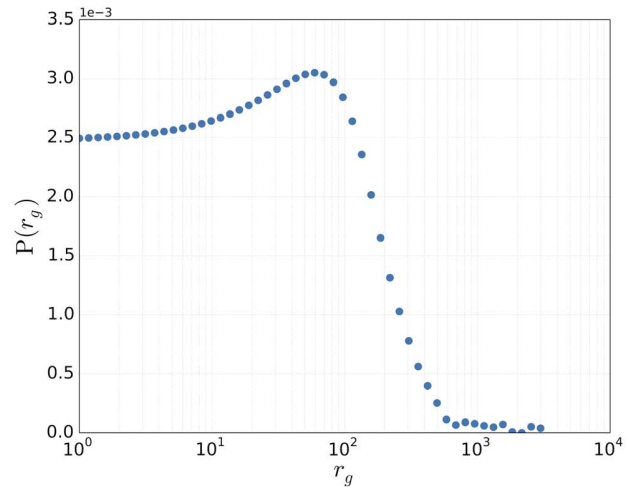
The interactions between the participants exhibit both daily and weekly rhythms. Figure 1 shows that the topology of the network of face-to-face meetings changes significantly within single day, revealing academic and social patterns formed by the students. Similarly, the intensity of the interactions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

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

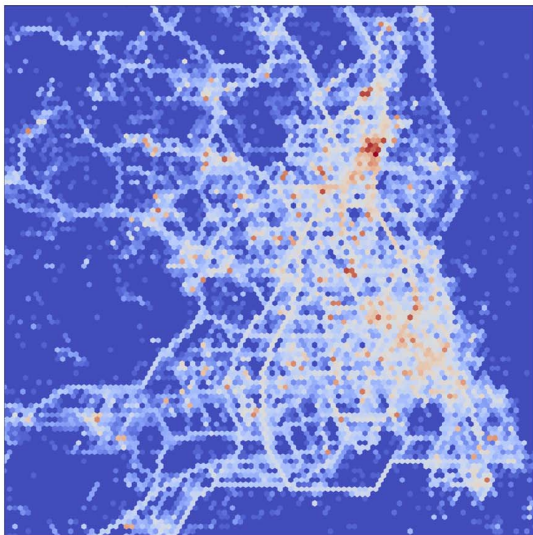
A estimated cumulative distribution function for the accuracy of samples a (in meters)



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



C histogram of the locations, with hex binning and logarithmic color scale



D cumulated transitions between stop locations

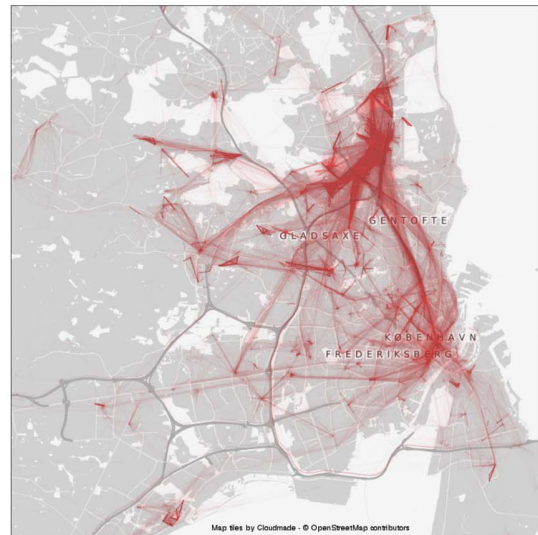


Figure 8. Location and Mobility. We show the accuracy of the collected samples, radius of gyration of the participants, and identify patterns of collective mobility.

doi:10.1371/journal.pone.0095978.g008

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

WiFi as an Additional Channel for Social Ties

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

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

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

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

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

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

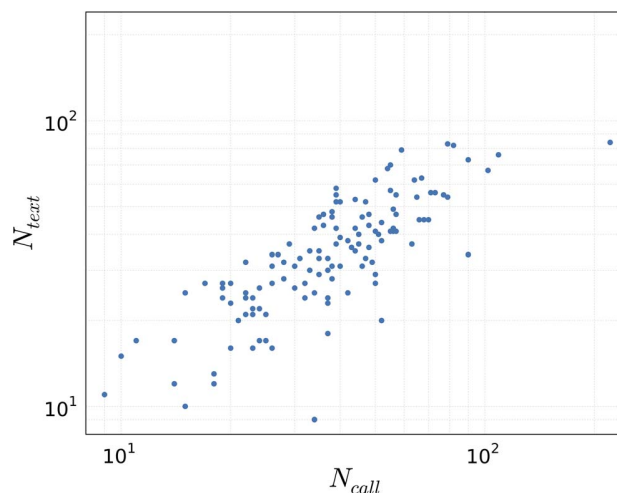


Figure 9. Diversity of communication logs. Diversity is estimated as the set of unique numbers that a person has contacted or been contacted by in the given time period on a given channel. We note a strong correlation in diversity (Pearson correlation of 0.75, $p < 0.05$), whereas the similarity of the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

Next, we compare the received signal strengths between overlapping routers using the mean ℓ_1 -norm (mean Manhattan distance, $\frac{\|X \cap Y\|_1}{|X \cap Y|}$). Received signal strength (RSSI) is measured

in *dBm* and the Manhattan distance between two routers is the difference in the RSSI between them, measured in *dB*. Thus, the mean Manhattan distance is the mean difference in received signal strength of the overlapping routers in the two compared scans.

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

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

Location and Mobility

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

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

We calculate the radius of gyration r_g as defined in [38] and approximate the probability distribution function using a gaussian kernel density estimation, see Figure 8B. We select the appropriate kernel bandwidth through leave-one-out cross-validation scheme from Statsmodels KDEMultivariate class [161]. The kernel density peaks around 10^2 km and then rapidly goes down, displaying a fat-tailed distribution. Manual inspection of the few participants with r_g around 10^3 km revealed that travels abroad can amount to

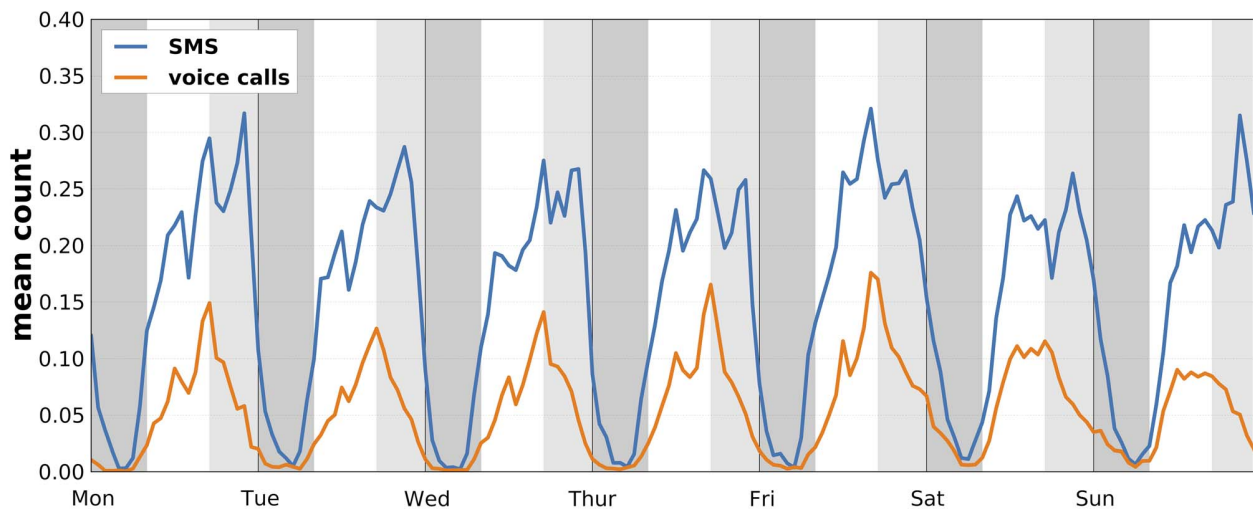


Figure 10. Weekly temporal dynamics of interactions. All calls and SMS, both incoming and outgoing, were calculated over the entire dataset and averaged per participant and per week, showing the mean number of interactions participants had in a given weekly bin. Light gray denotes 5pm, the time when lectures end at the university, dark gray covers night between 12 midnight and 8am. SMS is used more for communication outside regular business hours.

doi:10.1371/journal.pone.0095978.g010

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

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

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

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

Call and Text Communication Patterns

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

The ability to efficiently map communication networks and mobility patterns (using cell towers) for large populations has made it possible to quantify human mobility patterns, including investigations of social structure evolution [166], economic development [67], human mobility [37,38], spreading patterns [57], and collective behavior with respect to emergencies [60]. In

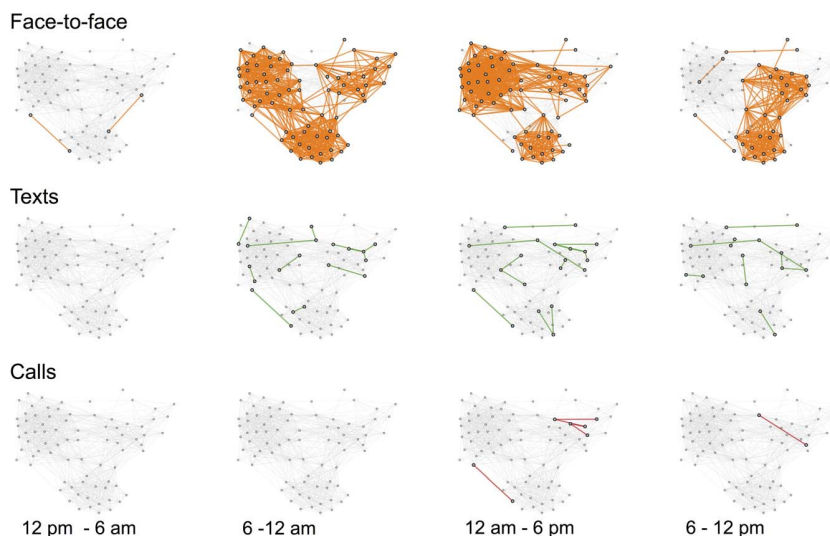


Figure 11. Daily activations in three networks. One day (Friday) in a network showing how different views are produced by observing different channels.

doi:10.1371/journal.pone.0095978.g011

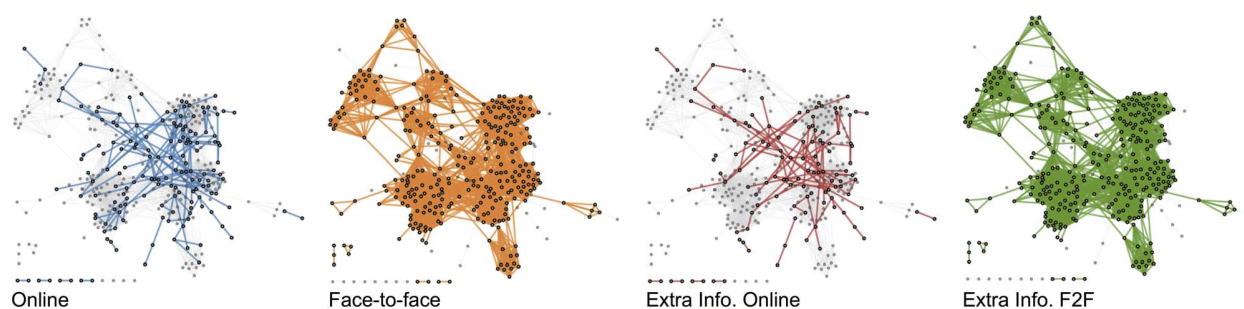
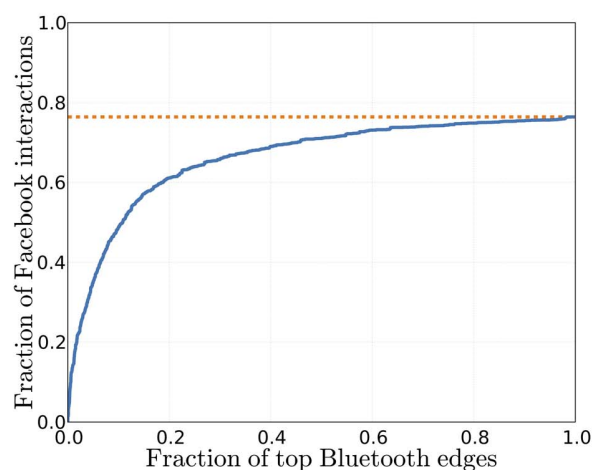
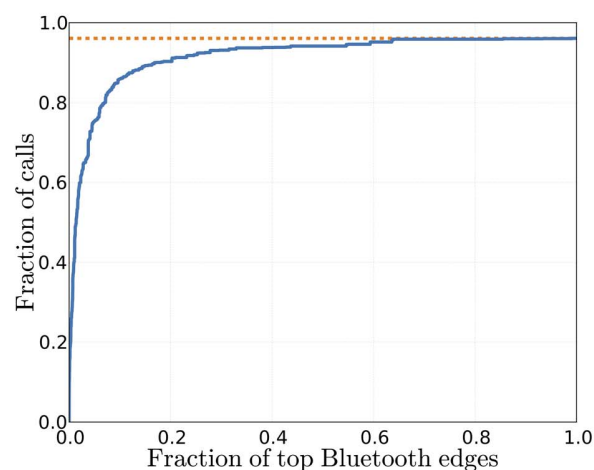


Figure 12. Face-to-face and online activity. The figure shows data from the 2013 deployment for one representative week. **Online:** Interactions (messages, wall posts, photos, etc.) between participants on Facebook. **Face-to-Face:** Only the most active edges, which account for 80% of all traffic,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

A face-to-face and Facebook interactions



B face-to-face and voice calls



C voice calls and Facebook interactions

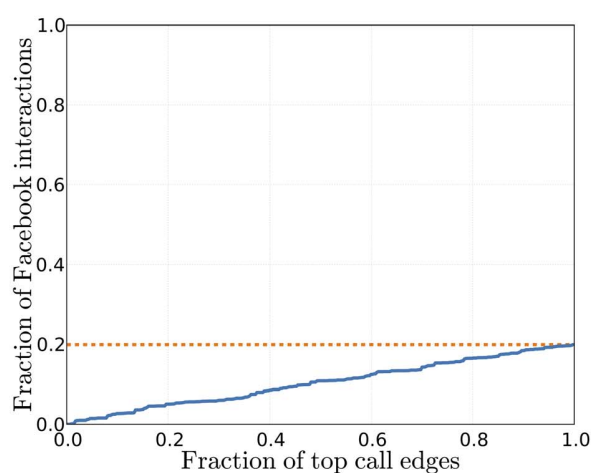


Figure 13. Network similarity. Defined as the fraction of ties from one communication channel that can be recovered by considering the top k fraction of edges from a different channel. Orange dashed line indicates the maximum fraction of ties the network accounts for. The strongest 10% of face-to-face interactions account for >50% of online ties and ~90% of call ties, while 23.58% of Facebook ties and 3.85% of call ties are not contained in the Bluetooth data. Between call and Facebook, the 10% strongest call ties account for <3% while in total >80% of Facebook ties are unaccounted. All values are calculated for interactions that took place in January 2014.
doi:10.1371/journal.pone.0095978.g013

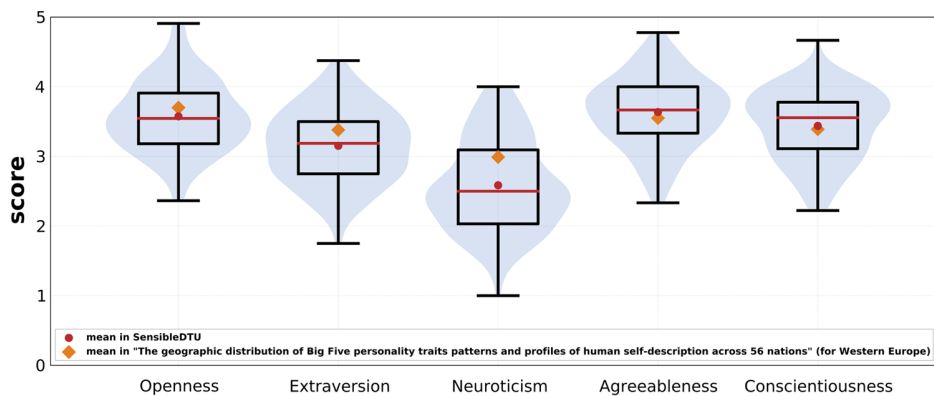


Figure 14. Personality traits. Violin plot of personality traits. Summary statistics are: **openness** $\mu_O = 3.58$, $\sigma_O = 0.52$; **extraversion** $\mu_E = 3.15$, $\sigma_E = 0.53$; **neuroticism** $\mu_N = 2.59$, $\sigma_N = 0.65$; **agreeableness** $\mu_A = 3.64$, $\sigma_A = 0.51$; **conscientiousness** $\mu_C = 3.44$, $\sigma_C = 0.51$. Mean values from our deployment (red circles) compared with mean values reported for Western Europe (mixed student and general population) [170] (orange diamonds). doi:10.1371/journal.pone.0095978.g014

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

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

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

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

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

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

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

Online friendships

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

A simple method for quantifying the similarity between two networks is to consider the fraction of links we can recover from them. Sorting face-to-face edges according to activity (highest first) we consider the fraction of online ties the top k Bluetooth links correspond to. Figure 13A shows that 10% of the strongest Bluetooth ties account for more than 50% of the Facebook interactions. However, as noted before, the Bluetooth channel does not recover all online interactions—23.58% of Facebook ties are unaccounted for. Applying this measure between Bluetooth

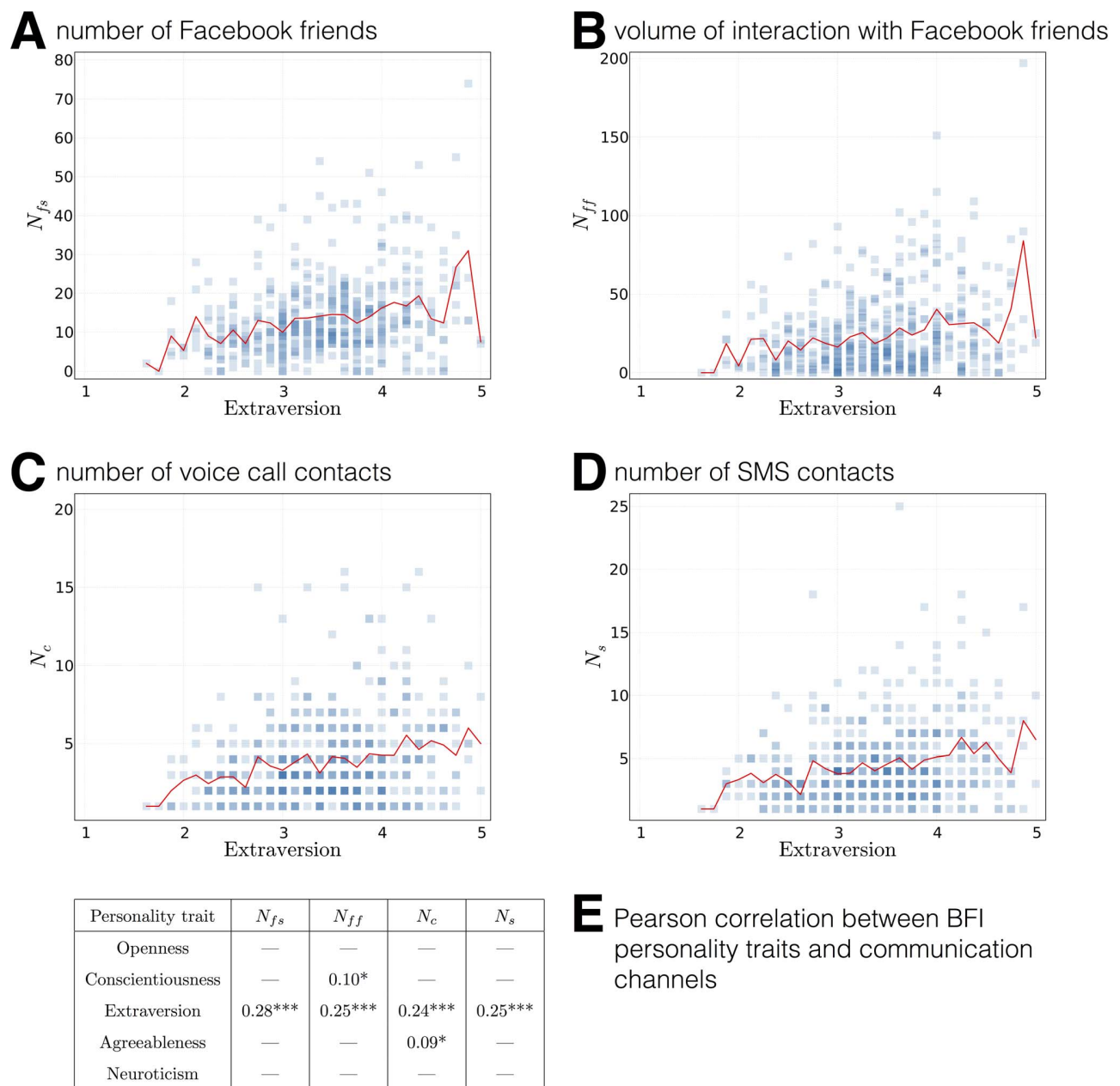


Figure 15. Correlation between personality traits and communication. Data from the 2013 deployment for $N=488$ participants, showing communication only with other study participants. Extraversion, the only significant feature across all networks is plotted. The red line indicates mean value within personality trait. Random spikes are due to small number of participants with extreme values. **E**) Pearson correlation between Big Five Inventory personality traits and number of Facebook friends N_{fs} , volume of interactions with these friends N_{ff} , number of friends contacted via voice calls N_c and via SMS N_s . *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. doi:10.1371/journal.pone.0095978.g015

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

Personality traits

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

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

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

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

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

Perspectives

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

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

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

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

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

Acknowledgments

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

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

References

- Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, et al. (2008) Detecting influenza epidemics using search engine query data. *Nature* 457: 1012–1014.
- Aral S, Walker D (2012) Identifying influential and susceptible members of social networks. *Science* 337: 337–341.
- Stopczynski A, Pietri R, Pentland A, Lazer D, Lehmann S (2014) Privacy in Sensor-Driven Human Data Collection: A Guide for Practitioners. *arXiv preprint arXiv:14035299*.
- Onnela JP, Saramäki J, Hyvönen J, Szabó G, Lazer D, et al. (2007) Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences* 104: 7332–7336.
- Cha M, Haddadi H, Benevenuto F, Gummadi PK (2010) Measuring user influence in Twitter: The million follower fallacy. *ICWSM* 10: 10–17.
- Shalizi CR, Thomas AC (2011) Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research* 40: 211–239.
- Barabási AL, Jeong H, Néda Z, Ravasz E, Schubert A, et al. (2002) Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications* 311: 590–614.
- Kossinets G, Watts DJ (2006) Empirical analysis of an evolving social network. *Science* 311: 88–90.
- Lazer D, Pentland A, Adamic L, Aral S, Barabási A, et al. (2009) Life in the network: the coming age of computational social science. *Science* (New York, NY) 323: 721.
- Wesołowski A, Eagle N, Noor AM, Snow RW, Buckee CO (2013) The impact of biases in mobile phone ownership on estimates of human mobility. *Journal of The Royal Society Interface* 10: 20120986.
- Madrigal A (2013) Dark social: We have the whole history of the web wrong. *The Atlantic*.
- Cattuto C, Van den Broeck W, Barrat A, Colizza V, Pinton JF, et al. (2010) Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLoS one* 5: e11596.
- Wu L, Weber B, Aral S, Brynjolfsson E, Pentland A (2008) Mining face-to-face interaction networks using sociometric badges: Predicting productivity in an IT configuration task. Available at SSRN 1130251.
- Polastre J, Szewczyk R, Culler D (2005) Telos: enabling ultra-low power wireless research. In: *Information Processing in Sensor Networks*, 2005. IPSN 2005. Fourth International Symposium on. IEEE, pp. 364–369.
- Raento M, Oulasvirta A, Eagle N (2009) Smartphones an emerging tool for social scientists. *Sociological methods & research* 37: 426–454.
- Chronis I, Madan A, Pentland AS (2009) SocialCircuits: the art of using mobile phones for modeling personal interactions. In: *Proceedings of the ICMI-MIMI'09 Workshop on Multimodal Sensor-Based Systems and Mobile Phones for Social Computing*. ACM, p. 1.
- Pentland AS (2008) *Honest Signals: how they shape our world*. MIT Press.
- Olguin D, Madan A, Cebrian M, Pentland A (2011) Mobile sensing technologies and computational methods for collective intelligence. *Next Generation Data Technologies for Collective Computational Intelligence*: 575–597.
- Miller G (2012) The smartphone psychology manifesto. *Perspectives on Psychological Science* 7: 221–237.
- Raento M, Oulasvirta A, Petit R, Toivonen H (2005) ContextPhone: A prototyping platform for context-aware mobile applications. *Pervasive Computing*. IEEE 4: 51–59.
- Mulder I, Ter Hofte G, Kort J (2005) SocioXensor: Measuring user behaviour and user experience in context with mobile devices. In: *Proceedings of Measuring Behavior*. pp. 355–358.
- Froehlich J, Chen MY, Consolvo S, Harrison B, Landay JA (2007) MyExperience: a system for in situ tracing and capturing of user feedback on mobile phones. In: *Proceedings of the 5th international conference on Mobile systems, applications and services*. ACM, pp. 57–70.
- Cornelius C, Kapadia A, Kotz D, Peebles D, Shin M, et al. (2008) AnonySense: privacy-aware people-centric sensing. In: *Proceedings of the 6th international conference on Mobile systems, applications, and services*. ACM, pp. 211–224.
- Miluzzo E, Lane N, Fodor K, Peterson R, Lu H, et al. (2008) Sensing meets mobile social networks: the design, implementation and evaluation of the CenceMe application. In: *Proceedings of the 6th ACM conference on Embedded network sensor systems*. ACM, pp. 337–350.
- Kostakos V, O'Neill E (2008) Cityware: Urban computing to bridge online and real-world social networks. *Handbook of research on urban informatics: The practice and promise of the real-time city*: 195–204.
- Miluzzo E, Cornelius C, Ramaswamy A, Choudhury T, Liu Z, et al. (2010) Darwin phones: the evolution of sensing and inference on mobile phones. In: *Proceedings of the 8th international conference on Mobile systems, applications, and services*. ACM, pp. 5–20.
- Hu X, Leung VC, Chu TH, Chan HC (2013) Vita: A crowdsensing-oriented mobile cyber-physical system. *IEEE Transactions on Emerging Topics in Computing* 1: 148–165.
- Larsen JE, Jensen K (2009) *Mobile Context Toolbox*. In: *Smart Sensing and Context*, Springer. pp. 193–206.
- Eagle N, Pentland A (2006) Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing* 10: 255–268.
- (2013). Funf Open Sensing Framework. URL <http://funf.org/>. [Online; accessed 19-March-2014].
- Aharony N, Pan W, Ip C, Khayal I, Pentland A (2011) Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing* 7: 643–659.
- Kiukkonen N, Blom J, Dousse O, Gatica-Perez D, Laurila J (2010) Towards rich mobile phone datasets: Lausanne data collection campaign. *Proc ICPS*, Berlin.
- Laurila J, Gatica-Perez D, Aad I, Blom J, Bornet O, et al. (2012) The mobile data challenge: Big data for mobile computing research. In: *Mobile Data Challenge by Nokia Workshop*, in conjunction with Int. Conf. on Pervasive Computing, Newcastle, UK.
- Olguin D, Weber B, Kim T, Mohan A, Ara K, et al. (2009) Sensible organizations: Technology and methodology for automatically measuring organizational behavior. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 39: 43–55.
- Karikoski J, Nelimarkka M (2011) Measuring social relations with multiple datasets. *International Journal of Social Computing and Cyber-Physical Systems* 1: 98–113.
- Cranshaw J, Toch E, Hong J, Kittur A, Sadeh N (2010) Bridging the gap between physical location and online social networks. In: *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*. ACM, pp. 119–128.
- González MC, Hidalgo CA, Barabási AL (2008) Understanding individual human mobility patterns. *Nature* 453: 779–782.
- Song C, Qu Z, Blumm N, Barabási AL (2010) Limits of predictability in human mobility. *Science* 327: 1018–1021.
- Sevtuk A, Ratti C (2010) Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks. *Journal of Urban Technology* 17: 41–60.
- Bagrow JP, Lin YR (2012) Mesoscopic structure and social aspects of human mobility. *PLoS One* 7: e37676.
- De Domenico M, Lima A, Musolesi M (2013) Interdependence and predictability of human mobility and social interactions. *Pervasive and Mobile Computing* 9: 798–807.
- Eagle N, Pentland AS, Lazer D (2009) Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences* 106: 15274–15278.
- Eagle N, Pentland AS (2009) Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology* 63: 1057–1066.
- Clauset A, Eagle N (2007) Persistence and periodicity in a dynamic proximity network. In: *DyDAn Workshop on Computational Methods for Dynamic Interaction Networks*.
- Onnela JP, Saramäki J, Hyvönen J, Szabó G, De Menezes MA, et al. (2007) Analysis of a largescale weighted network of one-to-one human communication. *New Journal of Physics* 9: 179.
- Granovetter MS (1973) The strength of weak ties. *American Journal of Sociology*: 1360–1380.
- Lambiotte R, Blondel VD, de Kerchove C, Huens E, Prieur C, et al. (2008) Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications* 387: 5317–5325.
- Onnela JP, Arbesman S, González MC, Barabási AL, Christakis NA (2011) Geographic constraints on social network groups. *PLoS One* 6: e16939.
- Hidalgo CA, Rodriguez-Sickert C (2008) The dynamics of a mobile phone network. *Physica A: Statistical Mechanics and its Applications* 387: 3017–3024.
- Miritello G, Lara R, Cebrian M, Moro E (2013) Limited communication capacity unveils strategies for human interaction. *Scientific reports* 3.
- Miritello G, Moro E, Lara R, Martínez-López R, Belchamber J, et al. (2013) Time as a limited resource: Communication strategy in mobile phone networks. *Social Networks* 35: 89–95.
- Sun L, Axhausen KW, Lee DH, Huang X (2013) Understanding metropolitan patterns of daily encounters. *Proceedings of the National Academy of Sciences* 110: 13774–13779.
- Kapoor A, Eagle N, Horvitz E (2010) People, quakes, and communications: Inferences from call dynamics about a seismic event and its influences on a population. In: *AAAI Spring Symposium: Artificial Intelligence for Development*.

Author Contributions

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

54. Madan A, Cebrian M, Moturu S, Farrahi K, Pentland S (2012) Sensing the health state of a community. *IEEE Pervasive Computing* 11: 36–45.
55. Salathé M, Kazandjieva M, Lee JW, Lewis P, Feldman MW, et al. (2010) A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences* 107: 22020–22025.
56. Madan A, Farrahi K, Gatica-Perez D, Pentland A (2011) Pervasive sensing to model political opinions in face-to-face networks. In: Lyons K, Hightower J, Huang E, editors, *Pervasive Computing*, Springer Berlin Heidelberg, volume 6696 of *Lecture Notes in Computer Science*. pp.214–231. doi: 10.1007/978-3-642-21726-5_14. URL http://dx.doi.org/10.1007/978-3-642-21726-5_14.
57. Wang P, González MC, Hidalgo CA, Barabási AL (2009) Understanding the spreading patterns of mobile phone viruses. *Science* 324: 1071–1076.
58. Isella L, Romano M, Barrat A, Cattuto C, Colizza V, et al. (2011) Close encounters in a pediatric ward: measuring face-to-face proximity and mixing patterns with wearable sensors. *PLoS One* 6: e17144.
59. Stehlé J, Voirin N, Barrat A, Cattuto C, Isella L, et al. (2011) High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS One* 6: e23176.
60. Bagrow JP, Wang D, Barabási AL (2011) Collective response of human populations to large-scale emergencies. *PLoS One* 6: e17680.
61. Karsai M, Perra N, Vespignani A (2013) The emergence and role of strong ties in time-varying communication networks. *arXiv preprint arXiv:13035966*.
62. Christakis NA, Fowler JH (2007) The spread of obesity in a large social network over 32 years. *New England Journal of Medicine* 357: 370–379.
63. Christakis NA, Fowler JH (2008) The collective dynamics of smoking in a large social network. *New England Journal of Medicine* 358: 2249–2258.
64. Lyons R (2011) The spread of evidence-poor medicine via awed social-network analysis. *Statistics, Politics, and Policy* 2.
65. Steglich C, Snijders TA, Pearson M (2010) Dynamic networks and behavior: Separating selection from influence. *Sociological Methodology* 40: 329–393.
66. Aral S, Muchnik L, Sundararajan A (2009) Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences* 106: 21544–21549.
67. Eagle N, Macy M, Claxton R (2010) Network diversity and economic development. *Science* 328: 1029–1031.
68. Blondel V, Krings G, Thomas I (2010) Regions and borders of mobile telephony in Belgium and in the Brussels metropolitan zone. *Brussels Studies* 42.
69. Mahato H, Kern D, Holleis P, Schmidt A (2008) Implicit personalization of public environments using Bluetooth. In: *CHI'08 extended abstracts on Human factors in computing systems*. ACM, pp. 3093–3098.
70. Klasnja P, Consolvo S, Choudhury T, Beckwith R, Hightower J (2009) Exploring privacy concerns about personal sensing. In: *Proceedings of the 7th International Conference on Pervasive Computing*. Berlin, Heidelberg: Springer-Verlag, *Pervasive '09*, pp. 176–183. doi: 10.1007/978-3-642-01516-8_13. URL http://dx.doi.org/10.1007/978-3-642-01516-8_13.
71. Altshuler Y, Aharony N, Elovici Y, Pentland A, Cebrian M (2011) Stealing reality: when criminals become data scientists (or vice versa). *Security and Privacy in Social Networks*: 133–151.
72. Shokri R, Theodorakopoulos G, Le Boudec J, Hubaux J (2011) Quantifying location privacy. In: *Security and Privacy (SP), 2011 IEEE Symposium on*. IEEE, pp. 247–262.
73. Lane N, Xie J, Moscibroda T, Zhao F (2012) On the feasibility of user de-anonymization from shared mobile sensor data. In: *Proceedings of the Third International Workshop on Sensing Applications on Mobile Phones*. ACM, p. 3.
74. Srivatsa M, Hicks M (2012) De-anonymizing mobility traces: using social network as a side-channel. In: *Proceedings of the 2012 ACM conference on Computer and communications security*. ACM, pp. 628–637.
75. Mislove A, Viswanath B, Gummadi KP, Druschel P (2010) You are who you know: inferring user profiles in online social networks. In: *Proceedings of the third ACM International Conference on Web search and data mining*. ACM, pp. 251–260.
76. Zhou B, Pei J (2008) Preserving privacy in social networks against neighborhood attacks. In: *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*. IEEE, pp. 506–515.
77. Cheng J, Fu AWc, Liu J (2010) K-isomorphism: privacy preserving network publication against structural attacks. In: *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. ACM, pp. 459–470.
78. Li T, Li N (2009) On the tradeoff between privacy and utility in data publishing. In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 517–526.
79. Narayanan A, Shmatikov V (2008) Robust de-anonymization of large sparse datasets. In: *Security and Privacy, 2008. SP 2008. IEEE Symposium on*. IEEE, pp. 111–125.
80. Sweeney L (2000) Simple demographics often identify people uniquely. *Health (San Francisco)*: 1–34.
81. Barbaro M, Zeller T, Hansell S (2006) A face is exposed for AOL searcher no. 4417749. *New York Times* 9: 8For.
82. de Montjoye YA, Hidalgo CA, Verleysen M, Blondel VD (2013) Unique in the crowd: The privacy bounds of human mobility. *Scientific reports* 3.
83. Sweeney L (2002) k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10: 557–570.
84. Machanavajjhala A, Kifer D, Gehrke J, Venkatasubramanian M (2007) l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1: 3.
85. Li N, Li T, Venkatasubramanian S (2007) t-closeness: Privacy beyond k-anonymity and l-diversity. In: *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE, pp. 106–115.
86. Dinur I, Nissim K (2003) Revealing information while preserving privacy. In: *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, pp. 202–210.
87. Dwork C, Nissim K (2004) Privacy-preserving datamining on vertically partitioned databases. In: *Advances in Cryptology—CRYPTO 2004*. Springer, , pp. 134–138.
88. Blum A, Dwork C, McSherry F, Nissim K (2005) Practical privacy: the SuLQ framework. In: *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, pp. 128–138.
89. Dwork C, Kenthapadi K, McSherry F, Mironov I, Naor M (2006) Our data, ourselves: Privacy via distributed noise generation. *Advances in Cryptology—EUROCRYPT 2006*: 486–503.
90. Chawla S, Dwork C, McSherry F, Smith A, Wee H (2005) Toward privacy in public databases. *Theory of Cryptography*: 363–385.
91. Rivest RL, Adleman L, Dertouzos ML (1978) On data banks and privacy homomorphisms. *Foundations of secure computation* 4: 169–180.
92. Gentry C (2009) A fully homomorphic encryption scheme. Ph.D. thesis, Stanford University.
93. Tebaa M, El Hajji S (2012) Homomorphic encryption applied to the cloud computing security. In: *Proceedings of the World Congress on Engineering*. volume 1, pp. 4–6.
94. Nachrig M, Lauter K, Vaikuntanathan V (2011) Can homomorphic encryption be practical? In: *Proceedings of the 3rd ACM workshop on Cloud computing security workshop*. ACM, pp. 113–124.
95. Popa R, Balakrishnan H, Blumberg A (2009) VPriv: protecting privacy in location-based vehicular services. In: *Proceedings of the 18th conference on USENIX security symposium*. USENIX Association, pp. 335–350.
96. Molina A, Salajegheh M, Fu K (2009) HICCUPS: health information collaborative collection using privacy and security. In: *Proceedings of the first ACM workshop on Security and privacy in medical and home-care systems*. ACM, pp. 21–30.
97. Zdancewic SA (2002) Programming languages for information security. Ph.D. thesis, Cornell University.
98. Sfaxi L, Abdellatif T, Robbana R, Lakhnech Y (2010) Information flow control of component-based distributed systems. *Concurrency and Computation: Practice and Experience*.
99. Zeldovich N, Boyd-Wickizer S, Mazières D (2008) Securing distributed systems with information flow control. In: *Proceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation*. USENIX Association, pp. 293–308.
100. Mundada Y, Ramachandran A, Feamster N (2011) Silverline: data and network isolation for cloud services. In: *Proceedings of the 3rd USENIX conference on Hot topics in cloud computing*. USENIX Association, pp. 13–13.
101. Pappas V, Kemerlis V, Zavou A, Polychronakis M, Keromytis AD (2012) CloudFence: Enabling users to audit the use of their cloud-resident data.
102. Ganjali A, Lie D (2012) Auditing cloud administrators using information ow tracking. In: *Proceedings of the 2012 ACM Conference on Computer and Communications Security (CCS)*. pp. 79–84.
103. Bonch D, Lipton R (1996) A revocable backup system. In: *USENIX Security Symposium*. pp. 91–96.
104. Perlman R (2005) The ephemizer: Making data disappear. Technical report, Mountain View, CA, USA.
105. Perlman R (2005) File system design with assured delete. In: *Security in Storage Workshop, 2005. SISW'05. Third IEEE International*. IEEE, pp. 6–pp.
106. Geambasu R, Kohno T, Levy A, Levy HM (2009) Vanish: Increasing data privacy with self-destructing data. In: *Proc. of the 18th USENIX Security Symposium*. p. 56.
107. Agrawal R, Haas PJ, Kiernan J (2003) Watermarking relational data: framework, algorithms and analysis. *The VLDB journal* 12: 157–169.
108. Cox IJ, Miller ML, Bloom JA (2000) Watermarking applications and their properties. In: *Information Technology: Coding and Computing, 2000. Proceedings. International Conference on*. IEEE, pp. 6–10.
109. Cox IJ, Linnartz JP (1998) Some general methods for tampering with watermarks. *Selected Areas in Communications, IEEE Journal on* 16: 587–593.
110. Cattuto C, Van den Broeck W, Barrat A, Colizza V, Pinton JF, et al. (2010) Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLoS One* 5: e11596.
111. Larsen JE, Sapiezynski P, Stopczynski A, Mørup M, Theodorsen R (2013) Crowds, Bluetooth, and Rock'N'Roll: Understanding music festival participant behavior. In: *Proceedings of the 1st ACM International Workshop on Personal Data Meets Distributed Multimedia*. New York, NY, USA: ACM, PDM '13, pp. 11–18. doi:10.1145/2509352.2509399. URL <http://doi.acm.org/10.1145/2509352.2509399>.
112. Ranjan G, Zang H, Zhang ZL, Bolot J (2012) Are call detail records biased for sampling human mobility? *SIGMOBILE Mob Comput Commun Rev* 16: 33–44.

113. Isaacman S, Becker R, Cáceres R, Kobourov S, Martonosi M, et al. (2011) Identifying important places in peoples lives from cellular network data. In: *Pervasive Computing*, Springer. pp. 133–151.
114. Mucha P, Richardson T, Macon K, Porter M, Onnela JP (2010) Community structure in time-dependent, multiscale, and multiplex networks. *Science* 328: 876–878.
115. Szell M, Lambiotte R, Thurner S (2010) Multirelational organization of large-scale social networks. *Proceedings of the National Academy of Sciences USA* 107: 13636–13641.
116. Madden M, Lenhart A, Cortesi S, Gasser U, Duggan M, et al. (2013). *Teens, Social Media, and Privacy*. URL http://www.pewinternet.org/~media/Files/Reports/2013/PIP_TeensSocialMediaandPrivacy.pdf. [Online; accessed 19-March-2014].
117. Kossinets G (2006) Effects of missing data in social networks. *Social Networks* 28: 247–268.
118. Laumann E, Marsden P, Pinsky D (1983) The boundary specification problem in network analysis. Sage Publications. pp. 18–34.
119. Saramäki J, Leicht E, López E, Roberts SG, Reed-Tsochas F, et al. (2014) Persistence of social signatures in human communication. *Proceedings of the National Academy of Sciences* 111: 942–947.
120. Holme P, Saramäki J (2012) Temporal networks. *Physics reports* 519: 97–125.
121. Shalizi C, Thomas A (2011) Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research* 40: 211–239.
122. Fowler J, Christakis N (2008) Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham heart study. *British Medical Journal* 337: a2338.
123. Christakis N, Fowler J (2009) Connected: the surprising power of our social networks and how they shape our lives. Little, Brown and Company.
124. Li I, Dey A, Forlizzi J (2010) A stage-based model of personal informatics systems. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, pp. 557–566.
125. Larsen JE, Cuttone A, Jørgensen SL (2013) QS Spiral: Visualizing periodic quantified self data. In: *CHI 2013 Workshop on Personal Informatics in the Wild: Hacking Habits for Health & Happiness*.
126. Cuttone A, Lehmann S, Larsen JE (2013) A mobile personal informatics system with interactive visualizations of mobility and social interactions. In: *Proceedings of the 1st ACM international workshop on Personal data meets distributed multimedia*. ACM, pp. 27–30.
127. Rocha L, Liljeros F, Holme P (2011) Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. *PLoS Computational Biology* 7: e1001109.
128. Lee S, Rocha LE, Liljeros F, Holme P (2012) Exploiting temporal network structures of human interaction to effectively immunize populations. *PLoS One* 7: e36439.
129. Fortunato S (2010) Community detection in graphs. *Physics Reports* 486: 75–174.
130. Gulbahce N, Lehmann S (2008) The art of community detection. *BioEssays* 30: 934–938.
131. Ahn YY, Bagrow JP, Lehmann S (2010) Link communities reveal multiscale complexity in networks. *Nature* 466: 761–764.
132. Fiedler M (1975) A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. *Czechoslovak Mathematical Journal* 25: 619.
133. Bagrow JP, Lehmann S, Ahn YY (2011). Robustness and modular structure in networks. *arxiv/1102.5085*.
134. (2013). Facebook reports first quarter 2013 results. URL <http://investor.fb.com/releasedetail.cfm?ReleaseID=761090>. [Online; accessed 19-March-2014].
135. John OP, Srivastava S (1999) The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research* 2: 102–138.
136. Rosenberg M (1989) *Society and the adolescent self-image* (rev). Wesleyan University Press.
137. Back MD, Kufner AC, Dufner M, Gerlach TM, Rauthmann JF, et al. (2013) Narcissistic admiration and rivalry: Disentangling the bright and dark sides of narcissism. *Journal of Personality and Social Psychology* 105: 1013.
138. Diener E, Emmons RA, Larsen RJ, Griffin S (1985) The satisfaction with life scale. *Journal of personality assessment* 49: 71–75.
139. Rotter JB (1966) Generalized expectancies for internal versus external control of reinforcement. *Psychological monographs: General and applied* 80: 1.
140. Russell DW (1996) UCLA loneliness scale (version 3): Reliability, validity, and factor structure. *Journal of personality assessment* 66: 20–40.
141. Sherer M, Maddux JE, Mercandante B, Prentice-Dunn S, Jacobs B, et al. (1982) The self-efficacy scale: Construction and validation. *Psychological reports* 51: 663–671.
142. Cohen S, Kamarck T, Mermelstein R (1983) A global measure of perceived stress. *Journal of health and social behavior*: 385–396.
143. Bech P, Rasmussen NA, Olsen LR, Noerholm V, Abildgaard W (2001) The sensitivity and specificity of the major depression inventory, using the present state examination as the index of diagnostic validity. *Journal of affective disorders* 66: 159–164.
144. Lund R, Nielsen LS, Henriksen PW, Schmidt L, Avlund K, et al. (2014) Content validity and reliability of the Copenhagen Social Relations Questionnaire. *Journal of aging and health* 26: 128–150.
145. Watson D, Clark LA, Tellegen A (1988) Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology* 54: 1063.
146. Ellen RF, Firth R (1984) *Ethnographic research: A guide to general conduct*. Academic Press London.
147. de Montjoye YA, Wang SS, Pentland A, Anh DTT, Datta A, et al. (2012) On the trusted use of large-scale personal data. *IEEE Data Eng Bull* 35: 5–8.
148. Shampianer K, Mazar N, Ariely D (2007) Zero as a special price: The true value of free products. *Marketing Science* 26: 742–757.
149. (2013). SensibleDTU informed consent form (da). URL https://github.com/MIT-Model-Open-Data-and-Identity-System/SensibleData-Service/blob/production_sensibledu1k/sensible_data_service/documents/service_informed_consent_da.txt.
150. (2013). SensibleDTU informed consent form (en). URL https://github.com/MIT-Model-Open-Data-and-Identity-System/SensibleData-Service/blob/production_sensibledu1k/sensible_data_service/documents/service_informed_consent_en.txt.
151. Sekara V, Lehmann S (2014) Application of network properties and signal strength to identify face-to-face links in an electronic dataset. *arXiv preprint arXiv:14015836*.
152. Stopczynski A, Larsen JE, Lehmann S, Dynowski L, Fuentes M (2013) Participatory Bluetooth sensing: A method for acquiring spatio-temporal data about participant mobility and interactions at large scale events. In: *Pervasive Computing and Communications Workshops (PERCOM Workshops)*, 2013 IEEE International Conference on. IEEE, pp. 242–247.
153. Krings G, Karsai M, Bernhardsson S, Blondel VD, Saramäki J (2012) Effects of time window size and placement on the structure of an aggregated communication network. *EPJ Data Science* 1: 1–16.
154. Ribeiro B, Nicola P, Baronchelli A (2013) Quantifying the effect of temporal resolution on timevarying networks. *Scientific reports* 3.
155. Whitehead M, Phillips T, Page M, Molina M, Wood C (2012). *European mobile industry observatory 2011*. URL <http://www.gsm.com/publicpolicy/wp-content/uploads/2012/04/emofullwebfinal.pdf>. [Online; accessed 19-March-2014].
156. LaMarca A, Chawathe Y, Consolvo S, Hightower J, Smith I, et al. (2005) Place Lab: Device positioning using radio beacons in the wild. In: *Pervasive Computing*, Springer. pp. 116–133.
157. Kjergaard MB, Nurmi P (2012) Challenges for Social Sensing Using WiFi Signals. In: *Proceedings of the 1st ACM workshop on Mobile systems for computational social science*. New York, NY, USA: ACM, MCSS '12, pp. 17–21. doi:10.1145/2307863.2307869. URL <http://doi.acm.org/10.1145/2307863.2307869>.
158. Carlotto A, Parodi M, Bonamico C, Lavagetto F, Valla M (2008) Proximity classification for mobile devices using Wi-Fi environment similarity. In: *Proceedings of the first ACM international workshop on Mobile entity localization and tracking in GPS-less environments*. ACM, pp. 43–48.
159. Carreras I, Matic A, Saar P, Osmani V (2012) Comm2Sense: Detecting proximity through smartphones. In: *Pervasive Computing and Communications Workshops (PERCOM Workshops)*, 2012 IEEE International Conference on. IEEE, pp. 253–258.
160. Lin M, Hsu WJ (2013) Mining GPS data for mobility patterns: A survey. *Pervasive and Mobile Computing*.
161. (2013). *statsmodels.nonparametric.kde.kdeunivariate*. URL <http://statsmodels.sourceforge.net/devel/generated/statsmodels.nonparametric.kde.KDEUnivariate.html>. [Online; accessed 19-March-2014].
162. Hariharan R, Toyama K (2004) Project Lachesis: parsing and modeling location histories. In: *Geographic Information Science*, Springer. pp. 106–124.
163. Zheng Y, Zhang L, Xie X, Ma WY (2009) Mining interesting locations and travel sequences from GPS trajectories. In: *Proceedings of the 18th international conference on World wide web*. ACM, pp. 791–800.
164. Montoliu R, Gatica-Perez D (2010) Discovering human places of interest from multimodal mobile phone data. In: *Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia*. ACM, p. 12.
165. Zheng VW, Zheng Y, Xie X, Yang Q (2010) Collaborative location and activity recommendations with GPS history data. In: *Proceedings of the 19th international conference on World Wide Web*. ACM, pp. 1029–1038.
166. Palla G, Barabási AL, Vicsek T (2007) Quantifying social group evolution. *Nature* 446: 664–667.
167. (2013). “Facebook Reports Third Quarter 2013 Results”. URL <http://investor.fb.com/releasedetail.cfm?ReleaseID=802760>. [Online; accessed 19-March-2014].
168. McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: Homophily in social networks. *Annual review of sociology*: 415–444.
169. Blüml V, Kapusta ND, Doering S, Brähler E, Wagner B, et al. (2013) Personality factors and suicide risk in a representative sample of the german general population. *PLoS One* 8: e76646.
170. Schmitt DP, Allik J, McCrae RR, Benet-Martínez V (2007) The geographic distribution of big five personality traits patterns and profiles of human self-description across 56 nations. *Journal of Cross-Cultural Psychology* 38: 173–212.

171. Quercia D, Lambiotte R, Stillwell D, Kosinski M, Crowcroft J (2012) The personality of popular facebook users. In: Proceedings of the ACM 2012 conference on computer supported cooperative work. ACM, pp. 955–964.
172. Swickert RJ, Rosentreter CJ, Hittner JB, Mushrush JE (2002) Extraversion, social support processes, and stress. *Personality and Individual Differences* 32: 877–891.
173. Staiano J, Lepri B, Aharony N, Pianesi F, Sebe N, et al. (2012) Friends don't lie: inferring personality traits from social network structure. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing. ACM, pp. 321–330.
174. Chittaranjan G, Blom J, Gatica-Perez D (2011) Who's who with big-five: Analyzing and classifying personality traits with smartphones. In: *Wearable Computers (ISWC)*, 2011 15th Annual International Symposium on. IEEE, pp. 29–36.
175. Kalish Y, Robins G (2006) Psychological predispositions and network structure: The relationship between individual predispositions, structural holes and network closure. *Social Networks* 28: 56–84.
176. de Montjoye YA, Quoidbach J, Robic F, Pentland AS (2013) Predicting personality using novel mobile phone-based metrics. In: *Social Computing, Behavioral-Cultural Modeling and Prediction*, Springer. pp. 48–55.

The Strength of Friendship Ties in Proximity Sensor Data

Vedran Sekara^{1*}, Sune Lehmann^{1,2}

1 Cognitive Systems, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark, **2** Niels Bohr Institute, University of Copenhagen, Østerbro, Denmark



Abstract

Understanding how people interact and socialize is important in many contexts from disease control to urban planning. Datasets that capture this specific aspect of human life have increased in size and availability over the last few years. We have yet to understand, however, to what extent such electronic datasets may serve as a valid proxy for real life social interactions. For an observational dataset, gathered using mobile phones, we analyze the problem of identifying transient and non-important links, as well as how to highlight important social interactions. Applying the Bluetooth signal strength parameter to distinguish between observations, we demonstrate that weak links, compared to strong links, have a lower probability of being observed at later times, while such links—on average—also have lower link-weights and probability of sharing an online friendship. Further, the role of link-strength is investigated in relation to social network properties.

Citation: Sekara V, Lehmann S (2014) The Strength of Friendship Ties in Proximity Sensor Data. PLoS ONE 9(7): e100915. doi:10.1371/journal.pone.0100915

Editor: Christopher M. Danforth, University of Vermont, United States of America

Received: January 17, 2014; **Accepted:** May 31, 2014; **Published:** July 7, 2014

Copyright: © 2014 Sekara, Lehmann. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The authors were funded by a Young Investigator Grant from the Villum Foundation (High Resolution Networks, awarded to SL). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors confirm that one of the authors, Sune Lehmann, is a PLOS ONE Editorial Board member. This does not alter the authors' adherence to PLOS ONE Editorial policies and criteria.

* Email: vese@dtu.dk

Introduction

Recognizing genuine social connections is a central issue within multiple disciplines. When do connections happen? Where do they take place? And with whom is an individual connected? These questions are important when working to understand and design urban areas [1,2], studying close-contact spreading of infectious diseases [3–5], or organizing teams of knowledge workers [6–9]. In spite of their importance, measuring social ties in the real world can be difficult.

In classical social science the standard approach is to use self-reported data. This method, however, is only practical for relatively small groups and suffers from cognitive biases, errors of perception, and ambiguities [10]. Further, it has been shown that the ability to capture behavioral patterns via self-reported data is limited in many contexts [11]. A different approach for uncovering social behavior is to use digital records from emails and cell phone communication [12–19]. Although such analyses have improved our understanding of social ties, they have left many important questions unanswered—are electronic traces a valid proxy for real social connections? Eagle et al. [20] began to answer this question by including a spatial component as part of their data, using the short range ($\sim 10m$) Bluetooth sensor embedded in study participants' smartphones to measure physical proximity. Their results show that proximity data closely reflects social interactions in many cases. But since it is easy to think of examples where reciprocal Bluetooth detection does not correspond to social interaction (e.g. transient co-location in dining hall) the question remains, which observations correspond to actual social connections and which are just noise?

Multiple alternatives have been proposed to Bluetooth for sensor-driven measurement of social interactions, each with particular strengths and weaknesses [21–31]. For example, Radio

Frequency Identification (RFID) badges have short interaction ranges ($1–4m$) and measure only face-to-face interactions, thus solving many of the resolution problems posed by Bluetooth [30,31]. This approach, however, confines interactions to occur within specific areas covered by special radio receivers and requires participants to wear custom radio tags on their chests at all times—unlike Bluetooth which is ubiquitous across many types of modern electronic devices.

Our investigation digs into the role of Bluetooth signal strength, using a dataset obtained from applications running on the cell phones of 134 students at a large academic institution. Each phone records and sends data to researchers about call and text logs, Bluetooth devices in nearby proximity, WiFi hotspots in proximity, cell towers, GPS location, and battery usage [32]. In addition, we combine the data collected via the phones with online data, such as social graphs from Facebook for a majority of the participants. The study continuously gathers data, but in this paper we focus on Bluetooth proximity data gathered for 119 days during the academic year of 2012–2013. Specifically, we focus on the received signal strength parameter and propose a methodology that applies signal strength to distinguish between social and non-social interactions. We concentrate on the signal parameter because it is present in a majority of digitally recorded proximity datasets [30,32,33] and in addition, it also suggests a rough estimate for the distance between two devices. Applying the method on our data, we compare the findings to a null model and demonstrate how removing links with low signal strength influences network structure. Moreover, we use estimated link-weights and an online dataset to validate the friendship-quality of removed links.

Materials and Methods

Dataset

We distributed phones among students from four study lines (majors), where each major was chosen based on the fraction of students interested in participating in the project. This selection method yielded a coverage of >93% of students per study line, enabling us to capture a dense sample of the social interactions between subjects. Such high coverage of internal connections within a social group, with respect to the density of social interactions combined with the duration of observation, has not been achieved in earlier studies [20,30].

The data collector application installed on each phone follows a predefined scanning time table, which specifies the activation and duration of each probe. Proximity data is obtained by using the Bluetooth probe. Every 300 seconds each phone performs a Bluetooth scan that lasts 30 seconds. During the scan it registers all discoverable devices within its vicinity (5–10m) along with the associated received signal strength indicator (RSSI) [34]. Recorded proximity data is of the form (*i*, *j*, *t*, *s*), denoting that person *i* has observed *j* at time *t* with signal strength *s*. Only links between experiment participants are considered, comprising a dataset of 2 183 434 time ordered edges between 134 nodes, see Table 1 for more information. Data collection, anonymization, and storage was approved by the Danish Data Protection Agency, and complies with both local and EU regulations. Written informed consent was obtained via electronic means, where all invited participants digitally signed the form with their university credentials. Along with the mobile phone study we also collected Facebook graphs of the participants. Not all users donated their data since this was voluntary, however we obtained a user participation of ~88% (119 users and 1018 Facebook friendships). For the missing 12% of users, we assume they do not share any online friendships with the bulk of participants.

Identifying links

Independent of starting conditions, the scanning framework on one phone will drift out of sync with the framework on other phones after a certain amount of time, thus the phones will inevitably scan in a desynchronized manner. This desynchronization can mainly be attributed to: internal drift in the time-protocol of each phone, depletion of the battery,

and users manually turning phones off. To account for irregular scans, we divide time into windows (bins) of fixed width and aggregate the Bluetooth observations within each time-window into a weighted adjacency matrix. The complete adjacency matrix is then given by: $W = (W^{(\Delta t_1)}, W^{(\Delta t_2)}, \dots, W^{(\Delta t_n)})$, where each link is weighted by its signal strength and where Δt_i indicates window number *i*. These matrices generally assume a non-symmetric form, i.e. person *A* might observe *B* with signal strength *s* while person *B* observes *A* with strength *s'*, or not at all. The scanning frequency of the application sets a natural lower limit of the network resolution to 5 minutes. If we are interested in the social dynamics at a different temporal resolution we can aggregate the adjacency matrices and retain entries according to some heuristic (e.g. with the strongest signal). Depending on the level of description (monthly, weekly, daily, hourly, or every 5 minutes) the researcher must think carefully about the definition of a network connection. Frameworks for finding the best temporal resolution, so called *natural timescales* have for specific problems been investigated by Clauset and Eagle [35], and Sulo et al. [36]. In this paper, however, we are interested in the identification and removal of non-social proximity links, so aggregating multiple time-windows is not a concern here. Henceforth we solely work with 5 minutes time-bins.

The Bluetooth probe logs all discoverable devices within a sphere with a radius of 5–10 meters—walls and floor divisions reduce the radius, but the reduction in signal depends on the construction materials [37]. Blindly taking proximity observations as a ground truth for social interactions will introduce both false negative and false positive links in the social network. False negative links are typically induced by hardware errors beyond our control, thus we focus on identifying false positive links. We therefore propose to identify non-social or noisy proximity links via the signal strength parameter. The parameter can be thought of as a proxy for the relative distance between devices, since most people carry their phones on them, it in principle also suggests the separation distance between individuals.

Previous work has applied Bluetooth signals to estimate the position of individuals [38–41] but studies by Hay [42], and Hossein et al. [43] have revealed signal strength as an unsuitable candidate for accurately estimating location. However, the complexity of the problem can greatly be reduced by focusing on the relative distance between individuals rather than position. In theory, the transmitted power between two antennae is inversely proportional to the distance squared between them [44]. Reality is more complicated, due to noise and reflection caused by obstacles.

We use the ideal result as a reference while we perform empirical measurements to determine how signal strength depends on distance. Two devices are placed on the ground in a simulated classroom setting, where we are able to control the relative distance between them. The resulting measurements are plotted in Fig. 1A. As is evident from the figure, there is a large variance in the measured signal strength values for each fixed distance. However, as both phones exhibit the same variance we can exclude faulty hardware; further, environmental noise such as interference from other devices, or solar radiation can also be dismissed since there appear no daily patterns in the data. But we observe multiple bands or so-called modes onto which measurements collapse. Ladd et al. [33] noted a similar behavior for the received signal strength of WiFi connections, both are phenomena caused by non-Gaussian distributed noise. The empirical measurements form a foundation for understanding signal variance as

Table 1. Data overview.

	Total	Average pr. time-bin
Nodes (Users)	134	17.32
Edges (Dyads)	2 183 434	62.50
Time-bins	34 272	-
Average clustering	0.85	0.26
Average degree	103.51	2.41

Statistics showing the number of total (aggregated) and average values of network properties. Time-bins span five minutes and cover the entire 119 day period, including weekends and holidays. For the average values we only take active nodes into account, i.e. people that have observed another person or been observed themselves in that specific time-bin. Network properties are calculated for the full aggregated network and as averages over each temporal network slice.

doi:10.1371/journal.pone.0100915.t001

a function of distance, but they were performed in a controlled environment. In reality, there are a multitude of ways to carry a smartphone: some carry it around in a pocket, others in a bag. Liu and Striegel [45] investigated how these various scenarios influence the received signal strength—their results indicate only minor variations, hence we conclude that the general behavior is similar to the measurements shown in the figure. Further, social interactions are not only limited to office environments, so we have re-produced the experiment outdoors and in basement-like settings; the results are similar.

Bi-directional observations yield at most two observations per dyad per 5-minute time-bin, we can average over the measurements (Fig 1B), or take the maximal value (Fig 1C). Fig. 2 shows the distributions of signal strength for each respective distance. For raw data, Fig. 2A, we observe a localized zero-distance distribution while the 1, 2, and 3-m distributions overlap considerably. Averaging over values per time-bin smoothes out and compresses the distributions, but the bulk of the distributions still overlap (Fig. 2B). Taking only the maximal signal value into account separates the distributions more effectively (Fig. 2C). The reasoning behind choosing the maximal signal value is that phones are physically at different locations and we expect the distance to be maximally reflected in the distributions.

Thus, by thresholding observations on signal strength, we can filter out proximity links that are likely to be further away than a certain distance. By doing so we are able to emphasize links that are more probable of being genuine social interactions, while minimizing noise and filtering away non-social proximity links. From the behavioral data we count the number of appearances per dyad and assign the values as weights for each link. Link weights follow a heavy-tailed distribution, with a majority of pairs only observed a few times (low weights), a social behavior that has previously been observed by Onnela et al. [15]. Based on their weight we divide links into two categories: weak and strong. A link is defined as ‘weak’ if it has been observed (on average) less than once per day during the data collection period, remaining links are characterized as ‘strong’. An effective threshold should maximize the number of removed weak links, while minimizing the loss of strong links. Fig. 3 depicts the number of weak and strong links as a function of threshold value. We observe that, as we increase the threshold, the number of weak links decreases linearly, while the number of strong links remains roughly constant and then drops off suddenly. Taking into account both the maximum-value distance distributions (Fig. 2C) and link weights (Fig. 3), we choose the value (-80 dBm) that optimizes the ratio between strong and weak links. In a large majority of cases, this corresponds to

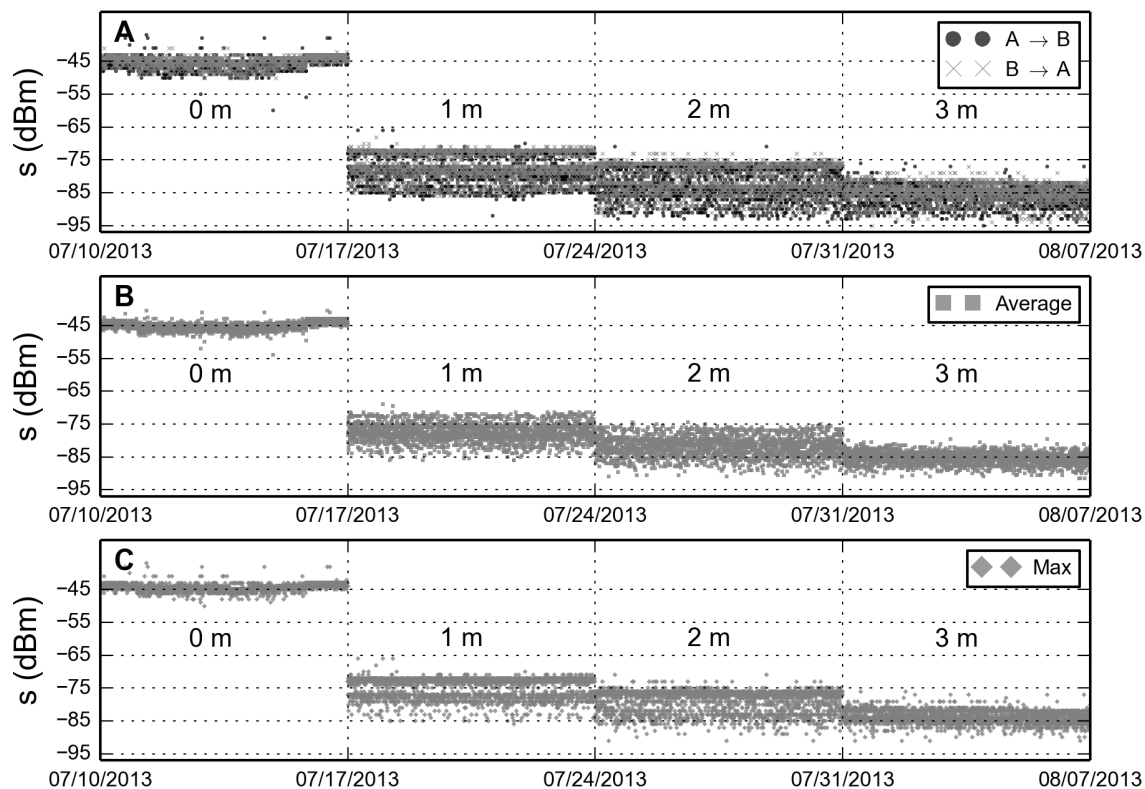


Figure 1. Bluetooth signal strength (RSSI) as a function of distance. **A:** Scans between two phones. Measurements are per distance performed every five minutes over the course of 7 days. Mean value and standard deviation per distance are respectively $\mu_0 = -45.13 \pm 1.56$ dBm, $\mu_1 = -77.48 \pm 4.15$ dBm, $\mu_2 = -82.03 \pm 4.57$ dBm, and $\mu_3 = -85.49 \pm 2.75$ dBm. **B:** Average of the values in respective time-bins. Summary statistics are: $\mu_0^{\text{avg}} = -45.13 \pm 1.20$ dBm, $\mu_1^{\text{avg}} = -77.46 \pm 2.90$ dBm, $\mu_2^{\text{avg}} = -81.99 \pm 3.17$ dBm, and $\mu_3^{\text{avg}} = -85.45 \pm 1.88$ dBm. **C:** Maximal value per time-bin. The mean value and standard deviation per distance are: $\mu_0^{\text{max}} = -44.41 \pm 1.11$ dBm, $\mu_1^{\text{max}} = -75.09 \pm 3.24$ dBm, $\mu_2^{\text{max}} = -79.25 \pm 3.47$ dBm, and $\mu_3^{\text{max}} = -83.88 \pm 2.00$ dBm. The measurements cover hypothetical situations where individuals are far from each other and on either side of a wall. doi:10.1371/journal.pone.0100915.g001

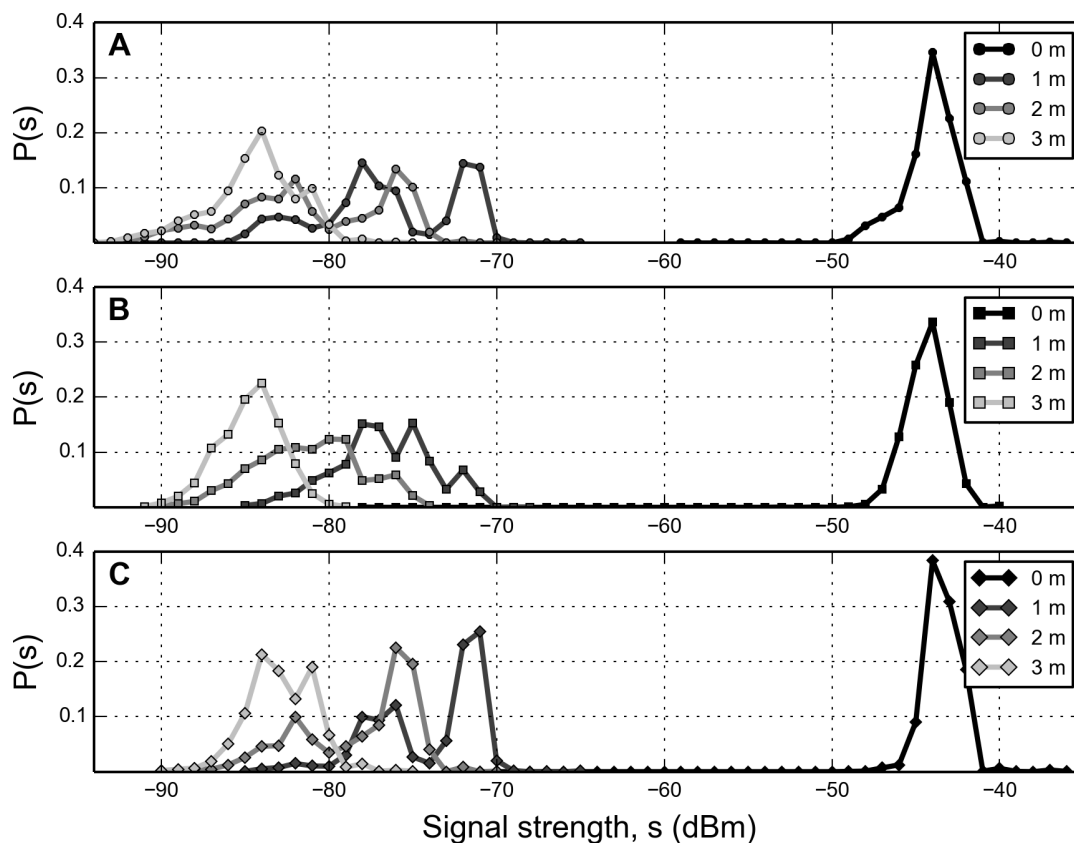


Figure 2. Distributions of signal strength for the respective distances. **A:** Raw data. Measurements from both phones are statistically indistinguishable and are collapsed into single distributions, i.e. there is no difference between whether *A* observes *B* or vice versa. **B:** Average of signal strength per time-bin. **C:** Maximal value of signal strength per time-bin.
doi:10.1371/journal.pone.0100915.g002

interactions that occur within a radius of 0–2 meters—a distance which Hall [46] notes as a typical social distance for interactions among close acquaintances.

Removing links

This section outlines various strategies for removing non-social links from the network. Fig. 4A shows an illustration of the raw proximity data for a single time-bin, a link is drawn if either $i \rightarrow j$ or $j \rightarrow i$. Thickness of a link represents the strength of the received signal. For the thresholded network (Fig. 4B) we remove links according to the strength of the signal (where we assume the weaker the signal the greater the relative distance between two persons). To estimate the effect of the threshold we compare it to a null model, where we remove the same number of links, but where the links are chosen at random, illustrated in Fig. 4C. To minimize any noise the random removal might cause, we repeat the procedure $n=100$ times, each time choosing a new set of random links, with statistics averaged over the 100 repetitions. As a reference, to check whether thresholding actually emphasizes social proximity links, we additionally compare it to a control network, where we remove the same amount of links, but where the links have signal strengths *above* or *equal* to the threshold, Fig. 4D. This procedure is also repeated n times. In a situation where there are more links below the threshold than above, we will remove fewer links for the latter compared to the other networks.

Results

Network properties

Now that we have determined a threshold for filtering out non-social proximity links, let us study the effects on the network properties. Thresholding weak links does not significantly influence the number of nodes present (N) in the network (Fig. 5A), while the number of links (M) is substantially reduced (Fig. 5B). On average we remove 2.38 nodes and 32.18 links per time-bin. Social networks differ topologically from other kinds of networks by having a larger than expected number of triangles [47], thus clustering is a key component in determining the effects of thresholding. Fig. 6 suggests that we are, in fact, keeping real social interactions: random removal disentangles the network and dramatically decreases the clustering coefficient, while thresholding conserves most of the average clustering. Calculating the average ratio ($\langle\langle c_T \rangle\rangle / \langle\langle c_N \rangle\rangle$) between clustering in the thresholded ($\langle\langle c_T \rangle\rangle$) and the null networks ($\langle\langle c_N \rangle\rangle$) reveals that c_T on average is 2.38 larger. These findings emphasize that a selection process based on signal strength greatly differs from a random one.

Link evaluation

Sorting links by signal strength and disregarding weak ones greatly reduces the number of links, but do we remove the correct links, i.e. do we get rid of noisy, non-social links? The fact that clustering remains high in spite of removing a large fraction of

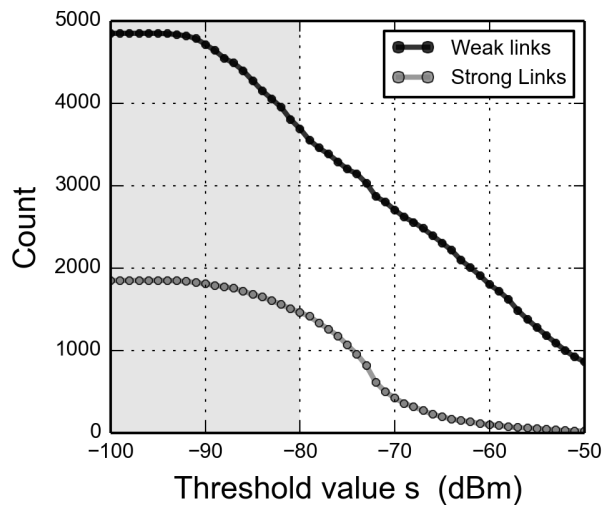


Figure 3. Number of links per type as a function of threshold value. Links are classified as weak if they are observed less than 120 times in the data, i.e. links that on average are observed less than once per day—otherwise they are classified as strong. Grouping students into study lines, reveals that links within each study line have an almost uniform distribution of weights while links across study lines are distributed according to a heavy-tailed distribution. A threshold of -80 dBm (gray area) removes 1159 weak and 387 strong links and classifies 97.6% of inter-study line links as weak and 86.7% of intra-study line links as strong.
doi:10.1371/journal.pone.0100915.g003

links is a good sign, but we want to investigate this question more directly. To do so, we divide the problem into two timescales; a short one where we consider the probability that a removed link might reappear a few time-steps later, and a long where we evaluate the quality of a removed link according to certain network properties. Let's first consider the short time-scale. We assume that human interactions take place on a time-scale that is mostly longer than the 5-minute time-bins we analyze here. Thus, if a noisy link is removed, the probability that it will re-appear in one of the immediately following time-steps should be low, since no interaction is assumed to take place. However, we expect the probability to be significantly greater than zero, since even weak (non-social) links imply physical proximity. Similarly, if we (accidentally) remove a social link, the probability that it will appear again should be high, since the social activity is expected to continue to take place.

Let us formalize this notion. Consider a link e that is removed at time t , the probability that the link will appear in the next time-

step is $p(t+1|e,t)$. Generalizing this we can write the probability that any removed link will appear in all the following n time-steps as:

$$p(t+1, \dots, t+n|t) = \frac{\text{no. links removed at } t \text{ present at } t+1 \cap \dots \cap t+n}{\text{no. links removed at } t} \quad (1)$$

Fig. 7A illustrates that thresholded links in subsequent time-steps are observed less frequently than both null and control links. To compare with the worst possible condition, we compare data from each thresholded time-bin with the *raw data* from the next bin (where the raw data contains many weak links). In spite of this, we observe a clear advantage of distinguishing between links with weak and strong signal strengths. If we look at values for $t+1$, the first subsequent time-step, the probability of re-occurrence in the thresholded network is about 12% lower than for the null model, and as we look to later time-steps, the gap widens.

A different set of social dynamics unfolds on longer timescales where the class schedule imposes certain links to appear periodically, e.g. every week. Here we determine impact of removing links in two ways. First, we use total link weights and second, we use online friendship status. Friends meet frequently; we capture this behavior by using the total number of observations of a certain dyad to estimate the weight of a friendship (again, counted in the raw network). Thus, we evaluate the quality of a removed links by considering its total weight compared to the weight of other links present in the same time-bin. However, since multiple links are removed per time-bin we are more interested in the average,

$$q_t = \frac{\text{Avg. weight of removed links at } t}{\text{Avg. weight of all links present at } t} \quad (2)$$

This estimates, per time-bin, whether removed links on average have weights below, close to, or above the mean. Note that the measure is intended to estimate the quality of removed links and is therefore not defined for bins where zero links are removed. Fig. 7B indicates difference in link selection processes. Choosing links at random (null network) removes both strong and weak links with equal probability, thus on average this corresponds to the mean weight of links present. Compared to null, the thresholded network removes links with weights below average, indicating that removed links are less frequently observed and therefore also less likely to be real friendships. The control case displays an diametrical behavior, on average, it removes links with higher weights.

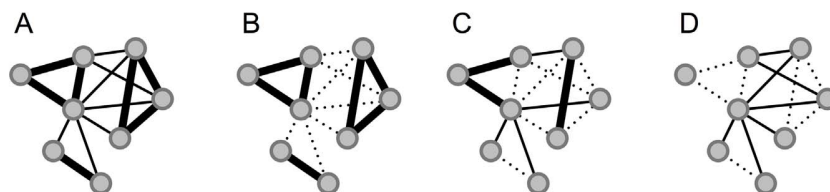


Figure 4. Networks. A: Raw network; shows all observed links for a specific time-bin. Thickness of a link symbolizes the maximum of the received signal strengths. B: Thresholded network, we remove links with received signal strengths below a certain threshold, where dotted lines indicate the removed links. C: Null model; with respect to the previous network we remove the same amount of links, but where the links are chosen at random. D: Control network, a similar amount of links with signal strength above or equal to the threshold are removed.
doi:10.1371/journal.pone.0100915.g004

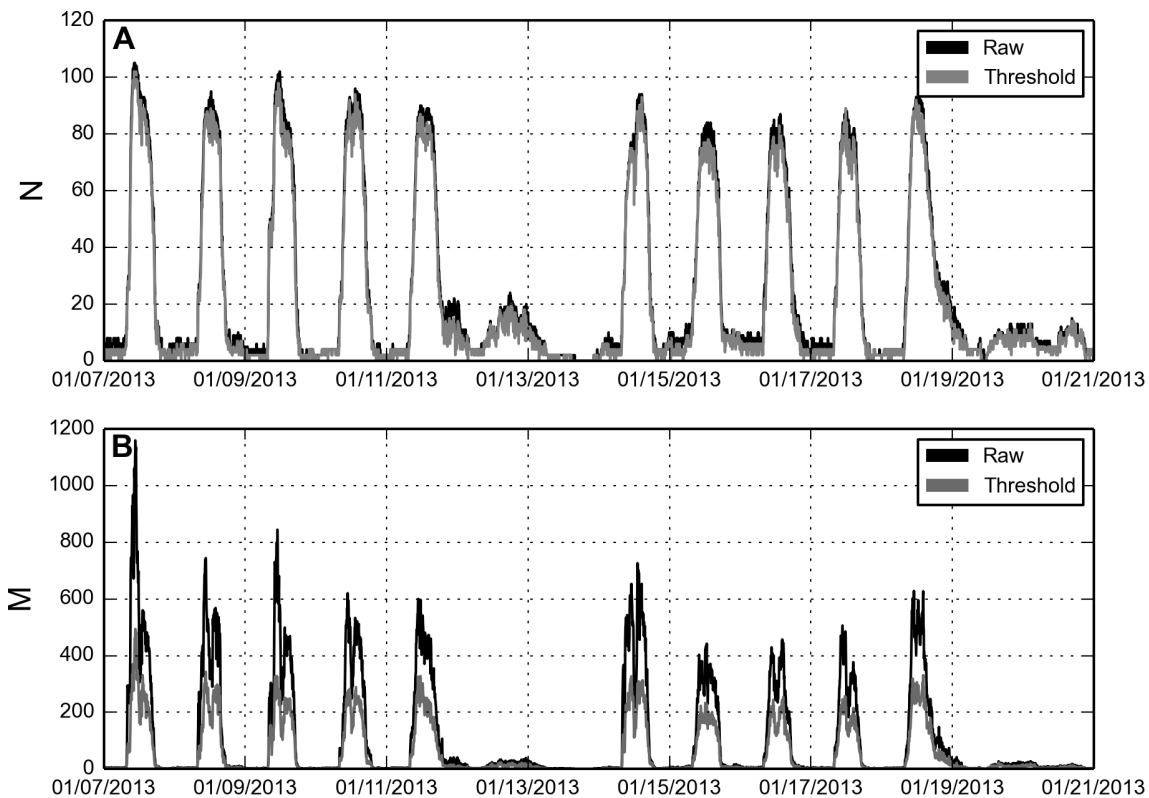


Figure 5. Network statistics. Properties are highly dynamic but on average we observe 17.32 nodes and 62.50 links per time-bin. **A:** Number of nodes N as a function of time. Only active nodes are counted, i.e. people that have observed another person or been observed themselves. Dynamics are shown for two weeks during the 2013 spring semester, clearly depicting both daily and weekly patterns. Data markers are omitted to avoid visual clutter. On average thresholding removes 3.06 nodes during weekends and holidays, and 2.38 during regular weekdays. **B:** Number of links M as a function of time. 10.60 links are on average removed during weekends/holidays, and 32.21 are removed during weekdays. doi:10.1371/journal.pone.0100915.g005

The second method to evaluate the link-selection processes compares the set of removed links with the structure of an online social network, i.e. if a removed proximity link has an equivalent online counterpart. We estimate the quality by measuring the fraction of removed links with respect to those present at time t .

$$q_t^{\text{FB}} = \frac{\text{no. of FB links removed at } t}{\text{no. of FB links present at } t} \quad (3)$$

The quality measure is essentially a ratio, i.e. it can assume values

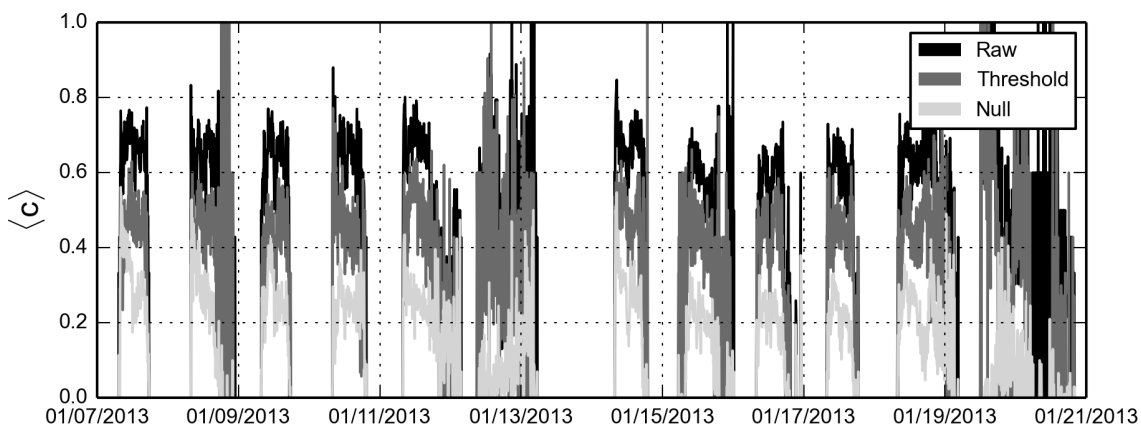


Figure 6. Average clustering. Only active nodes, i.e. nodes that are part of at least one dyad contribute to the average, the rest are disregarded. Average clustering is calculated according to the definition in [48]. Since social activity in groups larger than two individuals results in network triangles, the fact that clustering is not significantly reduced by thresholding (compared to the null model) provides evidence that we are preserving social structure in spite of link removal. doi:10.1371/journal.pone.0100915.g006

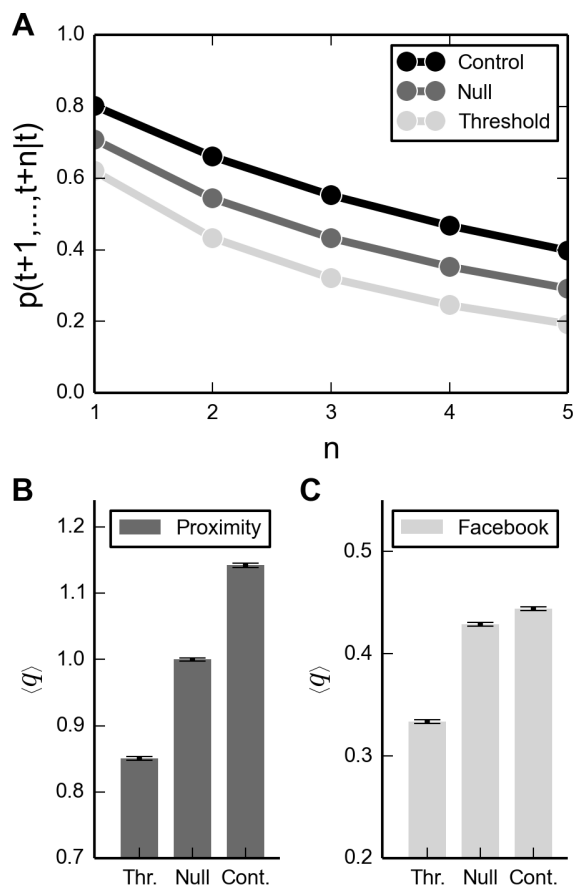


Figure 7. Link evaluation. **A:** Probability of link reappearance. For each selection process we remove a specific set of links. In the thresholded network, we remove links with weak signal strength. For the null network, we remove links at random. Lastly, in the control network case we remove strong links. The probability for links to reappear within all the next n time-steps is calculated using Eq. 1 and averaging over all time-bins. Boundary conditions are not applied and the reappearance probability for the last $n=5$ bins is not taken into account. **B:** Quality measure for proximity data. **C:** Quality measure for the online data. For each time-bin we calculate q_t as defined in Eq. 2 and 3. Brackets indicate a temporal average across all time-bins and value are shown for all three network types.
doi:10.1371/journal.pone.0100915.g007

$0 \leq q_t^{FB} \leq 1$ depending on the fraction of links that are removed. Bins with zero Facebook friendships are disregarded since they contain no information regarding the online social network. Fig. 7C shows that random removal (null network), on average, removes $\sim 43\%$ of online friendships, while the thresholded network removes $\sim 33\%$, a 10 percent point difference. For

comparison, the control network removes $\sim 44\%$ of the online links. Further, redoing the analysis for a dataset comprised only of users for which we have both proximity and online data for, does not significantly alter the results.

Facebook links are not necessary good indicators for strong friendships, but are more likely to correspond to real social interactions. In spite of this, both Fig. 7B and C support that distinguishing between strong and weak proximity links tends to emphasize real social interactions: on average thresholded links have lower edge weights and remove fewer Facebook friendships compared to both the null-model and the control.

Discussion

The availability of electronic datasets is increasing, so the question of how well can we use these electronic *clicks* to infer actual social interactions is important for effectively understanding processes such as relational dynamics, and contagion. Sorting links based on their signal strength allows us to distinguish between strong and weak ties, and we have argued that thresholding the network emphasizes social proximity links while eliminating some noise.

Simply thresholding links based on signal strength is not a perfect solution. In certain settings we remove real social connections while noisy links are retained. Our results indicate that the proposed framework is better at identifying strong links than removing them. A trend which the link-reappearance probability, link-weights, and online friendship analysis support. Compared to the baseline we achieve better results than just assuming all proximity observations as real social interactions. But determining whether a close proximity link corresponds to an actual friendship interaction is much more difficult. Multiple scenarios exist where people are in close contact but are not friends, one obvious example is queuing. Each human interaction has a specific social context, so an understanding of the underlying social fabric is required to fully discern when a close proximity link is an actual social meeting. This brings us back to the question of how to determine a real friendship from digital observations (cf. [10]). Close proximity may not be the best indicator of friendship; call logs, text logs, and geographical positions are all factors which coupled with information from the Bluetooth probe could give us a better insight into social dynamics and interactions.

Acknowledgments

We thank L. K. Hansen, A. Stopczynski, and P. Sapie ́ yński for many useful discussions and A. Cuttone for proofreading the manuscript.

Author Contributions

Conceived and designed the experiments: VS SL. Performed the experiments: VS SL. Analyzed the data: VS SL. Wrote the paper: VS SL.

References

1. Sun J, Yuan J, Wang Y, Si H, Shan X (2011) Exploring space-time structure of human mobility in urban space. *Physica A: Statistical Mechanics and its Applications* 390: 929–942.
2. Sevtsuk A, Ratti C (2010) Does urban mobility have a daily routine? learning from the aggregate data of mobile networks. *Journal of Urban Technology* 17: 41–60.
3. Liljeros F, Edling CR, Amaral LAN, Stanley HE, Åberg Y (2001) The web of human sexual contacts. *Nature* 411: 907–908.
4. Mossong J, Hens N, Jit M, Beutels P, Auranen K, et al. (2008) Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine* 5: e74.
5. Cauchemez S, Donnelly CA, Reed C, Ghani AC, Fraser C, et al. (2009) Household transmission of 2009 pandemic influenza A (H1N1) virus in the united states. *New England Journal of Medicine* 361: 2619–2627.
6. Wu L, Waber B, Aral S, Brynjolfsson E, Pentland A (2008) Mining face-to-face interaction networks using sociometric badges: Predicting productivity in an IT configuration task. Available at SSRN 1130251.
7. Pentland A (2012) The new science of building great teams. *Harvard Business Review* 90: 60–69.
8. Blansky D, Kavanaugh C, Boothroyd C, Benson B, Gallagher J, et al. (2013) Spread of academic success in a high school social network. *PLoS ONE* 8: e55944.

9. de Montjoye YA, Stopczynski A, Shmueli E, Pentland A, Lehmann S (2014) The strength of the strongest ties in collaborative problem solving. *Scientific reports* 4.
10. Wuchty S (2009) What is a social tie? *Proceedings of the National Academy of Sciences* 106: 15099–15100.
11. Watts DJ (2007) A twenty-first century science. *Nature* 445: 489–489.
12. Eckmann JP, Moses E, Sergi D (2004) Entropy of dialogues creates coherent structures in e-mail traffic. *Proceedings of the National Academy of Sciences of the United States of America* 101: 14333–14337.
13. Barabasi AL (2005) The origin of bursts and heavy tails in human dynamics. *Nature* 435: 207–211.
14. Kossinets G, Watts DJ (2006) Empirical analysis of an evolving social network. *Science* 311: 88–90.
15. Onnela JP, Saramäki J, Hyvönen J, Szabó G, Lazer D, et al. (2007) Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences* 104: 7332–7336.
16. Gonzalez MC, Hidalgo CA, Barabasi AL (2008) Understanding individual human mobility patterns. *Nature* 453: 779–782.
17. Lazer D, Pentland A, Adamic L, Aral S, Barabási AL, et al. (2009) Computational social science. *Science* 323: 721–723.
18. Song C, Qu Z, Blumm N, Barabási AL (2010) Limits of predictability in human mobility. *Science* 327: 1018–1021.
19. Bagrow JP, Wang D, Barabási AL (2011) Collective response of human populations to large-scale emergencies. *PLoS ONE* 6: e17680.
20. Eagle N, Pentland A, Lazer D (2009) Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences* 106: 15274–15278.
21. Haritaoglu I, Harwood D, Davis LS (2000) W4: Real-time surveillance of people and their activities. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22: 809–830.
22. Polastre J, Szewczyk R, Culler D (2005) Telos: enabling ultra-low power wireless research. In: *Information Processing in Sensor Networks, 2005. IPSN 2005. Fourth International Symposium on*. IEEE, pp. 364–369.
23. Salathé M, Kazandjieva M, Lee JW, Lewis P, Feldman MW, et al. (2010) A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences* 107: 22020–22025.
24. Rosenstein B (2008) Video use in social science research and program evaluation. *International Journal of Qualitative Methods* 1: 22–43.
25. Olguín DO, Waber BN, Kim T, Mohan A, Ara K, et al. (2009) Sensible organizations: Technology and methodology for automatically measuring organizational behavior. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 39: 43–55.
26. Kjærgaard MB, Nurmí P (2012) Challenges for social sensing using wifi signals. In: *Proceedings of the 1st ACM workshop on Mobile systems for computational social science*. ACM, pp. 17–21.
27. Wyatt D, Choudhury T, Kautz H (2007) Capturing spontaneous conversation and social dynamics: A privacy-sensitive data collection effort. In: *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. IEEE, volume 4, pp. IV–213.
28. Carreras I, Matic A, Saar P, Osmani V (2012) Comm2sense: Detecting proximity through smartphones. In: *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on*. IEEE, pp. 253–258.
29. Wang D, Pedreschi D, Song C, Giannotti F, Barabasi AL (2011) Human mobility, social ties, and link prediction. In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 1100–1108.
30. Cattuto C, Van den Broeck W, Barrat A, Colizza V, Pinton JF, et al. (2010) Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLoS ONE* 5: e11596.
31. Barrat A, Cattuto C (2013) Temporal networks of face-to-face human interactions. In: *Temporal Networks*, Springer. pp. 191–216.
32. Stopczynski A, Sekara V, Sapiezynski P, Cuttone A, Madsen MM, et al. (2014) Measuring largescale social networks with high resolution. *PLoS ONE* 9: e95978.
33. Ladd AM, Bekris KE, Rudys A, Kavraki LE, Wallach DS (2005) Robotics-based location sensing using wireless ethernet. *Wireless Networks* 11: 189–204.
34. Shorey R, Miller BA (2000) The bluetooth technology: merits and limitations. In: *Personal Wireless Communications, 2000 IEEE International Conference on*. IEEE, pp. 80–84.
35. Clauset A, Eagle N (2012) Persistence and periodicity in a dynamic proximity network. *arXiv preprint arXiv:12117343*.
36. Sulo R, Berger-Wolf T, Grossman R (2010) Meaningful selection of temporal resolution for dynamic networks. In: *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*. ACM, pp. 127–136.
37. Cheung KC, Intille SS, Larson K (2006) An inexpensive bluetooth-based indoor positioning hack. *Proc UbiComp06 Extended Abstracts*.
38. Anastasi G, Bandelloni R, Conti M, Delmastro F, Gregori E, et al. (2003) Experimenting an indoor bluetooth-based positioning service. In: *Distributed Computing Systems Workshops, 2003. Proceedings. 23rd International Conference on*. IEEE, pp. 480–483.
39. Bruno R, Delmastro F (2003) Design and analysis of a bluetooth-based indoor localization system. In: *Personal wireless communications*. Springer, pp. 711–725.
40. Madhavapeddy A, Tse A (2005) A study of bluetooth propagation using accurate indoor location mapping. In: *UbiComp 2005: Ubiquitous Computing*. Springer, pp. 105–122.
41. Zhou S, Pollard JK (2006) Position measurement using bluetooth. *Consumer Electronics, IEEE Transactions on* 52: 555–558.
42. Hay S, Harle R (2009) Bluetooth tracking without discoverability. In: *Location and context awareness*. Springer, pp. 120–137.
43. Hossain A, Soh WS (2007) A comprehensive study of bluetooth signal parameters for localization. In: *Personal, Indoor and Mobile Radio Communications, 2007. PIMRC 2007. IEEE 18th International Symposium on*. IEEE, pp. 1–5.
44. Friis HT (1946) A note on a simple transmission formula. *proc IRE* 34: 254–256.
45. Liu S, Striegel A (2011) Accurate extraction of face-to-face proximity using smartphones and bluetooth. In: *Computer Communications and Networks (ICCCN), 2011 Proceedings of 20th International Conference on*. IEEE, pp. 1–5.
46. Hall ET (1990) *The hidden dimension*. Anchor Books New York.
47. Newman ME, Park J (2003) Why social networks are different from other types of networks. *Physical Review E* 68: 036122.
48. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. *Nature* 393: 440–442.

Uncovering fundamental structures of dynamic social networks

Vedran Sekara¹, Arkadiusz Stopczynski^{1,2}, Sune Lehmann^{1,3*}

¹Department of Applied Mathematics and Computer Science,
Technical University of Denmark, Kgs. Lyngby, Denmark

²Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA

³The Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark

*To whom correspondence should be addressed; E-mail: sune.lehmann@gmail.com.

The minute-by-minute interactions within large, densely connected social systems are in a constant state of flux. While there has been impressive progress on understanding complex networks over the past decade, little is known about the regularities governing the micro-dynamics of such social networks. Here, we show that high-resolution data allow us to observe social gatherings directly. On the shortest time-scale, we find that gatherings are fluid, with members coming and going, but organized via a stable core of individuals. Cores exhibit a pattern of recurring meetings across weeks and months, each with varying degrees of regularity. Taken together, these findings provide a powerful simplification of the social system as a whole, resulting in a vocabulary for quantifying the complexity of social life. Using this theoretical framework, we demonstrate that, in analogy to human mobility, social behavior can be predicted with high precision.

Networks provide a powerful mathematical framework for analyzing the structure and dynamics of complex systems (1–3). The study of sociality among groups of humans has deep roots in the social science literature (4, 5) and community detection is a central component of modern network science, where communities have been found to be organized hierarchically as well as highly overlapping (6–9). Until this point, community detection in dynamic networks has required complex mathematical heuristics (7, 10, 11). Based on a unique dataset of social interactions across densely connected networks of physical proximity, telecommunication, as well as geolocation for a freshman class ($\sim 1\,000$ individuals) at a large university (12), we show that with high-resolution data describing social interactions, community detection becomes unnecessary. When time slices are shorter than the turnover rate, gatherings of individuals can be observed directly and without ambiguity (Fig. 1A-B). A simple matching across time slices then reveals the temporal development of each community (Fig. 1C and supporting material section S2).

Dynamically evolving gatherings represent a completely new object for quantitative study. Fig. 1C illustrates the dynamics. Each node is part of only a single gathering per time step, but may switch affiliations between co-existing gatherings. It is due to this gradual turnover that community detection has proven difficult in many other settings. Individuals participating in multiple groups create a highly overlapping structure, which is difficult to untangle (9) (Fig. 1A-B). The gatherings we discover, are broadly distributed in both size and duration, capturing meetings ranging from small cliques to large aggregations, and from short encounters on the order of minutes to prolonged interactions lasting many hours. We find that small groups tend to have shorter meetings while large groups have a typical duration of 1-2 hours. Since we are studying a population which works in the same physical location, we separate gatherings into *work* and *recreation*, where work-related gatherings may be driven by shared schedules. Partitioning gatherings according to location, we find 9 915 on-campus (work) and 13 872

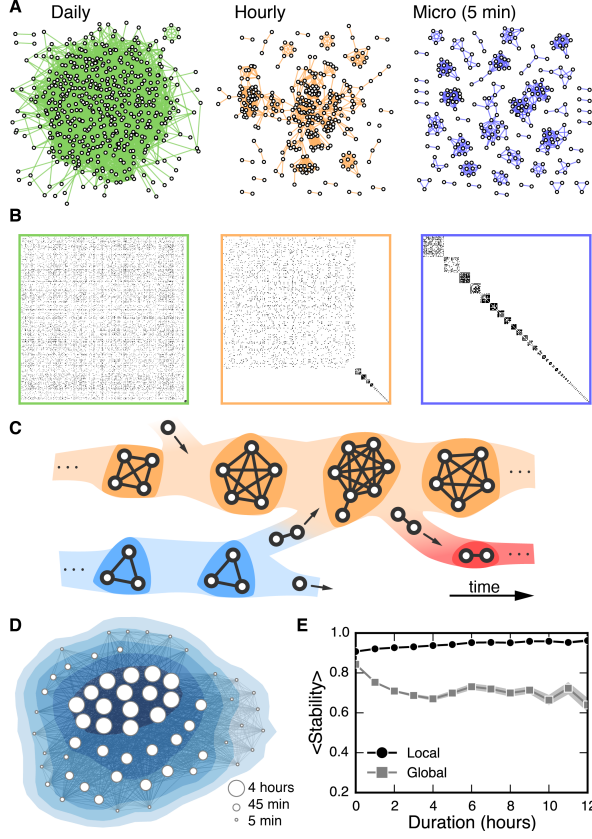


Figure 1: **Properties of gatherings.** (A) The network formed by physical proximity within one day (green), 60-minute (orange), and 5-minute temporal aggregation (blue). (B) Corresponding adjacency matrices sorted according to connected components. Groups are directly observable for short time-slices, but become overlapping as more time is aggregated in each bin. (C) Illustration of gathering dynamics. Gatherings change gradually with members flowing in and out of social contexts, participation in a gathering is given by at least one co-presence link. (D) Real world gatherings have soft boundaries, with nodes organized into a stable core with periphery nodes of lower participation levels. Node-size corresponds to participation. (E) The stability of gatherings as a function of duration. Global stability is defined as $\sum_{t_{\text{birth}}}^{t_{\text{death}}} J(g_t, G) / (t_{\text{death}} - t_{\text{birth}})$, where J denotes the Jaccard similarity and G is the aggregated network of slices ($G = g_{\text{birth}} \cup \dots \cup g_{\text{death}}$) while local stability is defined as $\sum_{t_{\text{birth}}}^{t_{\text{death}}-1} J(g_t, g_{t+1}) / (t_{\text{death}} - t_{\text{birth}} - 1)$.

off-campus (recreation) gatherings. Comparing work/recreation campus statistics, we find that recreational meetings tend to be smaller but last considerably longer, illustrating that the context of meetings can influence their properties (see supporting material sections S3 and S4). Unlike the typical community detection assumption of binary assignment to social contexts (13), we find that real world gatherings have soft boundaries (Fig. 1D), with some members participating for the total duration of the gathering, while others participate only briefly. We quantify this tendency in Fig. 1E where we investigate the stability of gatherings as a function of their duration. In terms of local stability (black line), which measures average turnover of nodes between subsequent network slices, we see that gatherings tend to be highly stable between time slices. When we compare each time slice to the aggregated network (global stability, gray line), we find that $\approx 70\%$ of all nodes are present in each slice. Both trends are largely independent of meeting duration. Comparing global to local stability, we see that high similarity between consecutive slices combined with a fixed global stability for any meeting duration implies the existence of stable cores. These cores consist of individuals who are present throughout the entirety of the meeting.

We now turn our attention to these stable cores and identify repeated appearances across the full duration of our dataset (section S4). We find a total of 7 320 such temporal communities. The number of appearances per core is a heavy tailed distribution; some cores appear only once, while the most active ones can appear multiple times per day over the full observation period (section S4.2). In the following, we focus on the temporal patterns of recurring gatherings, so we restrict our dataset to cores that, on average, are observed more than once per month. In analogy with gatherings, an important heterogeneity in our dataset is the split between *work cores*, which are mainly observed on campus, where meetings may be driven by externally imposed schedules and *recreation cores*, which are primarily observed elsewhere. Fig. 2A shows a clear difference between how individuals engage and spend time with respect to varying social

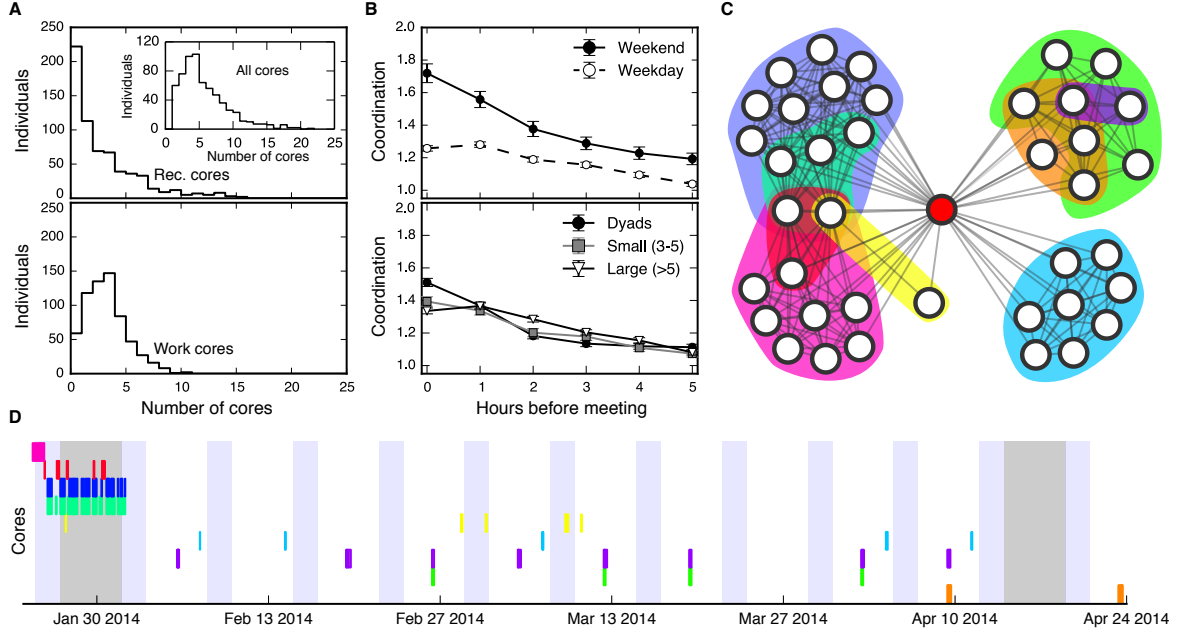


Figure 2: Cores summarize social contexts for individuals. (A) The distributions of work and recreational core membership, inset shows participation across both categories. Participation in recreational cores reveals that individuals typically participate in only one or two recreational contexts, although the tail of the distribution show some individuals with more gregarious behavior. The distribution of work cores is localized, with an average of 2.74 ± 1.85 work cores per individual, mainly reflecting participation in classes or group work. (B) Coordination prior to meetings, defined as $c_t = 1/N \sum_{n=1}^N a_t^n / \tilde{a}_t^n$, where a_t^n is the individual activity of person n in time-bin t , compared to an individual baseline denoted by the average activity \tilde{a}_t^n . More coordination is required to organize meetings during weekends than during weekdays, and larger meetings do not require additional coordination per participant. (C) Ego view of communities; we observe overlapping and hierarchically stacked structures. (D) The temporal complexity of participation for the cores in panel C. Time runs on the x -axis and each horizontal row of data corresponds to activation of a core. Gray and purple regions correspond to public holidays and weekends, respectively. We summarize information within this panel using time-correlated entropy (14).

contexts. Cores leave traces in other data channels that emphasize these differences. One such trace is coordination behavior, which we can explore by studying how call and text-message activity increases in the time leading up to a meeting. For each individual, we count the number of calls and texts within a hourly time-bin and compare to a null model based on typical hourly calling patterns for each participant. In this telecommunication network, we see clear evidence of coordination prior to meetings, which is accentuated during weekends when we expect meetings to be less schedule-driven (Fig. 2B). We also find that meetings require the same amount of prior communication per person regardless of the size of the gathering (Fig. 2B). In terms of network structure, we find that cores are highly overlapping, and large cores contain rich inner structure with hierarchically nested sub-cores (section S4.3). Here, however, we focus on cores from the perspective of individuals; Fig. 2C shows an ego-perspective. In Fig. 2D, we can observe the temporal patterns of core participation from late January to late April for the ego-network shown in Fig. 2C. The participation patterns are complex, displaying regularity mixed with randomness.

Cores provide a powerful simplification of the complexity of dynamic networks. A core represents a social context, and for an individual, the full set of cores provides a vocabulary for quantifying social life. With access to detailed mobility data Song *et al.* (14) made the highly surprising discovery that human mobility patterns contain great potential for predicting future locations based on past behavior. Below we show that—using high resolution data on social interactions encoded as cores—we are able to demonstrate how our social interactions allow for even higher levels of predictability than our mobility behavior. Given a sequence of social contexts, we can use the time-correlated entropy (14) to construct a measure of the complexity for each person in our dataset. In order to incorporate the full complexity of social encounters in the calculation, we include cores with any number of appearances as well as dyadic cores. The time-correlated entropy quantifies the amount of uncertainty within a data sequence, accounting

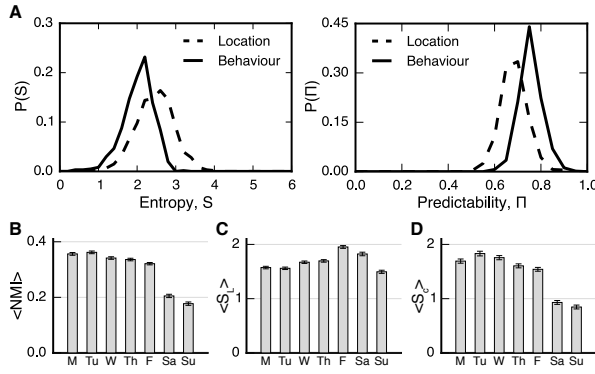


Figure 3: Geospatial and social predictability. (A) The distributions of entropy and predictability for social and location patterns. We find that overall social patterns tend towards lower entropy than geospatial traces, resulting in higher predictability. The fact that our location-predictability is lower than previously found (14) is connected to a number of factors. For example, our location data is based on GPS rather than cell towers and has significantly higher precision (15) (see section S5). (B) The average daily normalized mutual information between social and location sequences. Notice a significant drop on weekends. (C) The average daily entropy of location sequences, which increases on Friday and Saturdays, indicates increased geographical exploration on those days. (D) The average daily entropy of social engagements. The entropy is reduced on weekends, indicating a simpler pattern of social engagements, in agreement with Fig. 2A. In panels B-D we use the time-uncorrelated entropy to quantify the behavioral complexity.

for both for frequency and ordering of states and simultaneously provides an upper bound of the predictability based on routines in their social life. Fig. 3A shows the distribution of entropy and predictability. Social activity in our population is characterized by low temporal entropy, resulting in an average routine-based predictability limit of approximately 80%.

The core-representation allows us to examine existing results on predictability based on routine in location data (14) in the light of social patterns. Fig. 3A shows the distributions of entropy and predictability, calculated separately for the sequence of social states and spatial locations, respectively. Comparing social and location traces leads to a number of interesting findings. Firstly, we find that the social behavior tends to be more predictable and routine-driven than geospatial behavior. Secondly, the overall level of social and geospatial predictability is not correlated for individuals (p -value= 0.85 and section S5.3). Thus, highly routine driven location sequences do not imply predictable social behavior or *vice versa*. Thirdly, while the overall element of routine in a social or geospatial trajectory is not correlated for an individual, predictability in both contexts is closely related to daily and weekly schedules (16). During the week, our social and location behavior is correlated; we tend to meet the same people in the same places. This correlation between our social and location behavior is reduced on weekends. In Fig. 3B, we use the average (uncorrelated) normalized mutual information between daily social and geolocation traces to illustrate this behavior. The mutual information is a measure of how much knowing one variable reduces uncertainty about the other. Interestingly, we find that location-entropy increases during weekends, indicating a more exploratory behavior (Fig 3C). During that same time-period, social traces become simpler and more predictable (Fig 3D), consistent with Fig. 3A and previous work (17). Thus, in our population, periods of geospatial exploration are associated with social consolidation.

Up to now, prediction has implied understanding future behavior based on past routine. We now demonstrate how the social cores effectively summarize information in the underlying net-

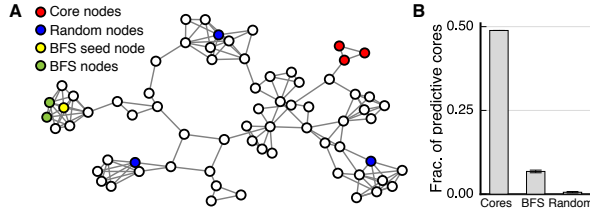


Figure 4: Features of social meetings. (A) Illustration of null models. Starting from co-located nodes in the daily graph, we select nodes (1) chosen randomly and (2) found using breadth first search (BFS) in a daily graph. (B) Social prediction using incomplete cores to predict arrival of remaining core-members. The strong increase compared to the BFS model emphasizes that full cores are needed for prediction, not just pairwise friendships. The ability to predict is tested on a month where cores have not been inferred. Error bars are calculated over $n = 100$ independent trials.

work by proposing a completely new kind of prediction, based on the cohesion of the social fabric itself. It is a well established fact that there is a correlation of spatial patterns between individuals that share a social tie (18–21). Pairs of traces, however, do not contain information that reveal at which times two location traces overlap, making it non-trivial to use this information for prediction. Cores provide such a temporal signature—an incomplete set of core members implies that the remaining members will arrive shortly. We illustrate this phenomenon on cores of size three. Given that two members of a core are observed, we measure the probability that the remaining member will arrive within one hour. To avoid testing on scheduled meetings, we only consider weekends and weekday evenings and nights (6pm-8am), where meetings are not driven by an academic schedule. Furthermore, we test on a month of data that has not been used for identifying cores. We now compare social prediction using cores to two null models (Fig. 4A). In the first, random, null model we create reference groups by randomly choosing groups of nodes from a daily graph. For the second null model, we form reference groups by performing a breadth first search (BFS) on the daily graph of interactions. But while Fig. 4B shows that $\sim 50\%$ of cores are predictive, both the random and the BFS reference groups fare poorly. By requiring that nodes share social connections, as well as a spatial location, the BFS

null model demonstrates that that it is not simply pairwise friendships that are predictive. The reason we are able to predict the arrival of final group member, is that the social context requires all core members to be present.

Within the existing literature, incorporating a temporal dimension dramatically complicates the mathematical description of complex networks (22). Here, we find the opposite. By observing social gatherings at the right time scale, when the temporal granularity is higher than the turnover rate, a simple matching across time slices reveals dynamically changing gatherings with stable cores that can be matched across time, providing a strong simplification of the social dynamics. These cores manifest in other data channels, such as through coordination behavior, and provide a finite vocabulary, which dramatically simplifies individuals' social activity. As a demonstration of the saliency of the description, we use the cores to a) quantify predictability within the social realm and b) allow for a new kind of non-routine prediction, based solely on the signal encoded in the core representation. Our work provides a first quantitative look at the rich patterns encoded in the micro-dynamics of a large system of closely interacting individuals, characterized by a high degree of order and predictability. The work presented here provides a new framework for describing human behavior and hints at the promise of our approach. We have focused on predictability, but we expect our work will support better modeling of processes in social systems, from epidemic modeling to urban planning, as well the science of team performance and public health.

References and Notes

1. D. Easley, J. Kleinberg, *Networks, crowds, and markets: Reasoning about a highly connected world* (Cambridge University Press, 2010).
2. M. Newman, *Networks: An Introduction* (Oxford University Press, 2010).

3. S. Wasserman, K. Faust, *Social Network Analysis: Methods and Applications* (Cambridge university press, 1994).
4. G. Simmel, *The Sociology of Georg Simmel* pp. 87–177 (1950).
5. E. Goffman, *Interaction ritual: Essays in Face to Face Behavior* (AldineTransaction, 2005).
6. G. Palla, I. Derényi, I. Farkas, T. Vicsek, *Nature* **435**, 814 (2005).
7. G. Palla, A.-L. Barabási, T. Vicsek, *Nature* **446**, 664 (2007).
8. A. Clauset, C. Moore, M. E. Newman, *Nature* **453**, 98 (2008).
9. Y.-Y. Ahn, J. P. Bagrow, S. Lehmann, *Nature* **466**, 761 (2010).
10. P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, J.-P. Onnela, *Science* **328**, 876 (2010).
11. L. Gauvin, A. Panisson, C. Cattuto, *PLoS ONE* **9**, e86028 (2014).
12. A. Stopczynski, *et al.*, *PLoS ONE* **9**, e95978 (2014).
13. S. Fortunato, *Physics Reports* **486**, 75 (2010).
14. C. Song, Z. Qu, N. Blumm, A.-L. Barabási, *Science* **327**, 1018 (2010).
15. M. Lin, W.-J. Hsu, Z. Q. Lee, *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (ACM, 2012), pp. 381–390.
16. J. McInerney, S. Stein, A. Rogers, N. R. Jennings, *Nokia Mobile Data Challenge Workshop* (2012).
17. N. Eagle, A. Pentland, *Personal and Ubiquitous Computing* **10**, 255 (2006).

18. D. J. Crandall, *et al.*, *Proceedings of the National Academy of Sciences* **107**, 22436 (2010).
19. D. Wang, D. Pedreschi, C. Song, F. Giannotti, A.-L. Barabási, *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2011), pp. 1100–1108.
20. E. Cho, S. A. Myers, J. Leskovec, *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2011), pp. 1082–1090.
21. M. De Domenico, A. Lima, M. Musolesi, *Pervasive and Mobile Computing* **9**, 798 (2013).
22. P. Holme, J. Saramäki, *Physics Reports* **519**, 97 (2012).
23. We thank L. K. Hansen, P. Sapiezynski, A. Cuttone, D. Wind, J. E. Larsen, B. S. Jensen, D. D. Lassen, M. A. Pedersen, A. Blok, T. B. Jørgensen, and Y. Y. Ahn for invaluable discussions and comments on the manuscript and R. Gatej for technical assistance. This work was supported a Young Investigator Grant from the Villum Foundation (High Resolution Networks, awarded to S.L.), and interdisciplinary UCPH 2016 grant (Social Fabric). Due to privacy implications we cannot share data but researchers are welcome to visit and work under our supervision.

Uncovering fundamental structures of
dynamic social networks
Supplementary Materials

Vedran Sekara, Arkadiusz Stopczynski & Sune Lehmann

March 16, 2015

Contents

S1 Summary of main results	4
S2 Data	5
S2.1 Construction of temporal network slices	5
S2.2 Selection of time-scales	6
S3 Gatherings	6
S3.1 Detecting gatherings	6
S3.1.1 Partitioning the dendrogram	7
S3.1.2 Temporal decay function	8
S3.1.3 Gathering timescales	9
S3.2 On- & off-campus gatherings	10
S3.3 Gathering statistics	11
S3.4 Temporal communities	13
S3.4.1 Optimal clustering partition	14
S3.5 Dyadic relations	16
S3.5.1 Dyad statistics	17
S4 Cores	17
S4.1 Extracting cores	17
S4.2 Core statistics	19
S4.3 Subcores	19
S4.4 Work & recreational cores	19
S4.5 Meeting regularity	21
S4.6 Ego viewpoint	21
S5 Predicting behavior from routine	21
S5.1 Comparing with previous studies	24
S5.1.1 Difference in populations	24
S5.1.2 Geospatial resolution	25
S5.1.3 Effects of binning	25
S5.2 Data for prediction	26
S5.2.1 Social vocabulary	26
S5.2.2 Location vocabulary	26
S5.2.3 Convergence of states	27
S5.3 Prediction	27
S5.4 Temporal aspect of predictability	28
S6 Social prediction	30
S6.1 Null models	30
S6.2 Comparison	31
S7 Coordination of meetings	31

List of Figures

S1	Summary of main findings	4
S2	Correlation between network slices	6
S3	Temporal coupling	7
S4	Illustration of group dendrogram	8
S5	Stability measures of gatherings	9
S6	Effect of decay parameter on robustness	10
S7	Gathering micro fluctuations	10
S8	Statistical features of gatherings	12
S9	Duration as function of gathering size	12
S10	Gathering stability versus size and duration	13
S11	Temporal patterns of gatherings	13
S12	Gathering participation profile	15
S13	Estimating optimal number of clusters using the Gap measure	16
S14	Statistical features of dyadic gatherings	17
S15	Extracting cores	18
S16	Core selection boundary	18
S17	Core statistics	19
S18	Number of cores per individual	20
S19	Subcore structures and statistics	20
S20	Distinction between work & recreational cores	22
S21	Meeting regularity of cores	23
S22	Ego-centric perspective of cores	23
S23	Cell tower resolution	25
S24	Effects of binning on predictability	26
S25	Time-series of states	27
S26	Distribution of the number of distinct states	27
S27	Distribution of entropy and predictability	28
S28	Correlation between social and location entropy	28
S29	Temporal aspects of predictability	29
S30	Relation between social and geolocation traces	30
S31	Social prediction	31
S32	Increased coordination prior to a meeting	32

List of Tables

S1	Data overview	5
----	-------------------------	---

S1 Summary of main results

Figure S1 illustrates our main findings. Social groups display a complex temporal behavior, with dynamics spanning multiple time-scales. Typically, incorporating the temporal dimension drastically complicates the mathematical description of complex networks such that community detection methods require sophisticated mathematical heuristics to disentangle the web of interactions. By observing social interactions at the right time scale—when the temporal granularity is higher than the turnover rate—we can directly observe social gatherings. Figure S1a-c shows social networks obtained using three temporal-windows of increasing size. While daily and hourly windows of aggregation obscure social relations (Fig. S1a-b), a micro-level description directly reveals the fundamental structures. Applying a simple mathematical matching scheme across time-slices reveals dynamically evolving gatherings with soft boundaries and stable cores (Fig. S1d). Unlike the typical community detection assumption of binary assignment, it is clear that some members participate for the total duration of the gathering, while others only participate briefly (Fig. S1d). Matching cores across longer time-scales allows us to observe dynamics that unfold over weeks and months. Cores provide a strong simplification of the social dynamics (Fig. S1e), and are manifested throughout other data channels such as coordination behavior via call and text messages. To demonstrate the saliency of our description we use the social contexts provided by cores to quantify the predictability of social life and give a proof of concept of a new type of non-routine prediction.

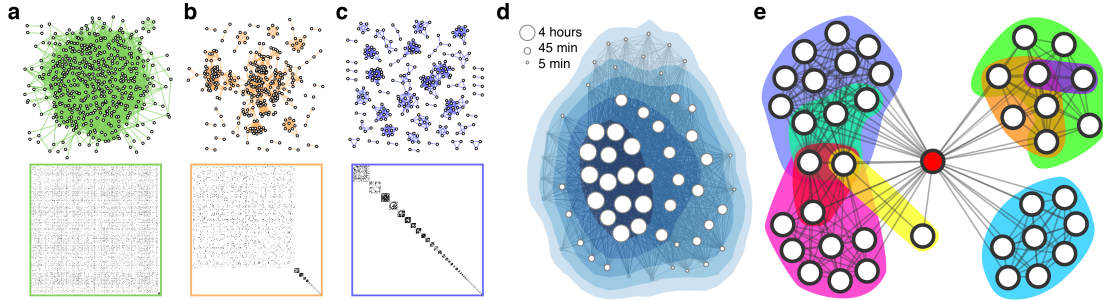


Figure S1: Summary of main findings. **a-c**, Network slices obtained by slicing the social dynamics using varying temporal windows (1 day (green), 1 hour (orange), and 5 minutes (blue)). Below, adjacency matrices colored in agreement with networks, and sorted according to component size. **d**, Gatherings have soft boundaries. The size of each node represents the level of participation. **e**, Cores simplify social dynamics and provide a context for social interactions.

The remainder of this document is organized as follows. Section 2 describes the dataset, Section 3 explains the details of how gatherings are constructed, and describes their basic statistics, including a discussion of dyads. Section 4 shows how cores are extracted and goes into detail on their temporal behavior, as well as the sub-core structures. In Section 5 we provide full details regarding the routine-based geospatial prediction as well as the social context prediction presented in the main text. Section 6 provides background on our model for purely social prediction and, finally, Section 7 discusses background information regarding the coordination leading up to meetings.

S2 Data

We consider a dataset from the *Copenhagen Networks Study*. It spans multiple years and measures with high resolution: physical interactions, telecommunications, online social networks, and geographical location. In addition, the dataset contains background information on all participants (personality, demographics, health, politics). These data are collected for a densely connected population of approximately 1 000 students at a large European university. Data is collected by running custom built applications installed on 1 000 smartphones (Google Nexus 4). Full details can be found in Ref. [1].

In this manuscript we focus on detecting and tracking co-located groups of individuals during a representative period of five months (roughly one semester), collected between January 1st and June 1st of 2014. The Bluetooth sensor collects proximity data ($\sim 0 - 10$ m) of the form (i, j, t, s) , where each interaction implies that person j has been in proximity of person i at time t , where the devices observe each other with signal strength s [2]. Bluetooth scans do not constitute a perfect proxy for face-to-face interactions. In fact multiple scenarios exist where people in close proximity do not interact and vice versa, nevertheless Bluetooth can successfully be applied in order to sense social networks [1–3]. Further, our gathering/core-description naturally filters our spurious connections by considering social structures that occur over across extended periods of time. Gatherings and cores are identified in the proximity network, and the remaining communication channels are used for validation purposes. In addition, we reserve proximity data from the month of May for validation purposes. Table S1 shows statistics across the various data sources for 814 individuals, on whom we focus due to their high data quality.

Data source	Total	Unique
Bluetooth interactions	14 673 869	154 818
Call & text interactions	75 364	1 216
Geographic locations	18 603 072	-
WiFi access points	1 663 483 977	2 412 702

Table S1: Data overview from January 1st – June 1st. Bluetooth and call & text logs are summarized for within-participant relations and do not include external interactions. The unique field denotes the number of distinct observation of each quantity, e.g. number of uniquely observed links.

S2.1 Construction of temporal network slices

The data collection application triggers Bluetooth scans every five minutes from the time a phone is turned on; for this reason, the collection of sensors does not follow a global schedule. To account for this behavior we divide all temporal information into absolute time-windows, Δ minutes wide. Within each temporal bin, we draw a unweighted undirected link between two individuals if either i has seen j or vice versa.

S2.2 Selection of time-scales

The scanning frequency of the application sets a natural lower limit of the network resolution to 5 minutes, however, there is no such upper limit for aggregation. So-called natural timescales have previously been investigated for specific networks with respect to their global topological properties [4–6]. Here we consider the correlation between slices (or turnover of nodes between slices) defined as $C = |E_i^\Delta \cap E_{i+1}^\Delta| / |E_i^\Delta \cup E_{i+1}^\Delta|$, where E_i^Δ denotes the set of edges that are observed in bin i with given temporal width Δ . According to Fig. S2 the average correlation decreases sharply as function of window-size, achieving maximum correlation for small bins. Slicing network dynamics into short slices (high resolution) also disentangles the network (Fig. S1), thus when time slices are shorter than the group’s turnover rate, we can directly observe individuals’ group affiliations. Based on Fig. S2 we chose a temporal width of 5 minutes, but windows of 10 minutes could have been chosen without deterioration of results.

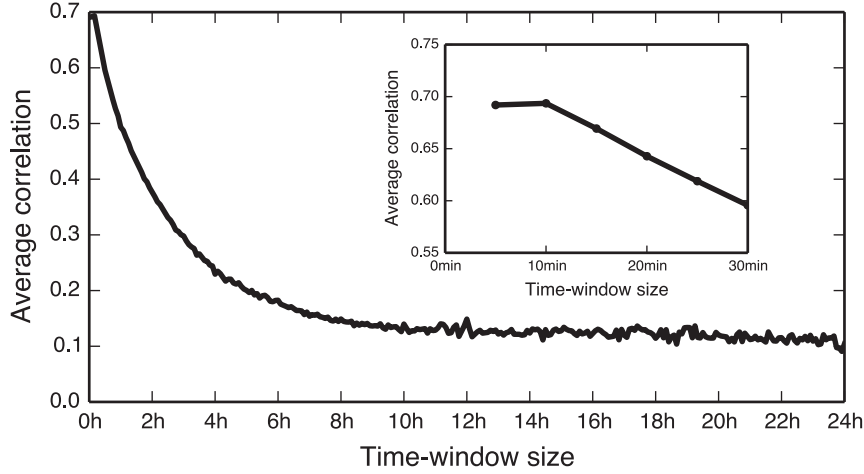


Figure S2: Average correlation between network slices, averaged across all time-bins. Inset shows a closeup for the smallest bin-sizes. Calculated for proximity data form March 2014.

S3 Gatherings

In this section we describe how connected components in the proximity network are matched across short timescales into dynamical ensembles, which we denote as gatherings. We then present fundamental statistics on gatherings (size distribution, durations, stability, start/end times), and analyze these properties in the light of on/off-campus behavior. Finally, we focus on identifying repeated gatherings across longer timescales to infer dynamical communities.

S3.1 Detecting gatherings

In each temporal slice we identify connected components, i.e. nodes that are in close physical proximity as social groups. Since dyadic relationships qualitatively differ from

group relations [7, 8] we generally treat components of size two separately.

A gathering is defined as a group that is persistent across time. To identify gatherings we apply agglomerative hierarchical matching, a widely used method, that merges groups based on their mutual distance (defined below) [9, 10]. Each group is initially assigned to its own cluster, then every iteration-step merges the two clusters with smallest distance according to the single linkage criteria ($\min(d_i(c_t, c_{t'}))$). This merge criterion is chosen because it is strictly local and will agglomerate clusters into chains, a preferable effect when clustering groups across time. The clustering process is repeated until all groups have been merged into a single cluster. Distance between groups is calculated using a modified version of the Jaccard similarity:

$$d(c_t, c_{t'}) = 1 - \frac{|c_t \cap c_{t'}|}{|c_t \cup c_{t'}|} f(\Delta t, \gamma), \quad (\text{S1})$$

where $f(\Delta t, \gamma)$ is a term that denotes the coupling between temporal slices and $\Delta t = t' - t$ denotes the temporal distance between two bins (for consecutive bins $\Delta t = 1$). The function can assume any form, increasing or decreasing; we utilize it to model decay between temporal slices, with the two most prominent forms being: exponential ($\exp(-\gamma(\Delta t - 1))$), and power-law ($\Delta t^{-\gamma}$), see Fig. S3. Thus, by definition the term assumes the value 1 (zero decay) between two consecutive temporal slices. For computational reasons we only focus on gatherings identified using exponential decay with $\gamma = 0.4$, other decay parameters yield similar results see SI Section S3.1.2.

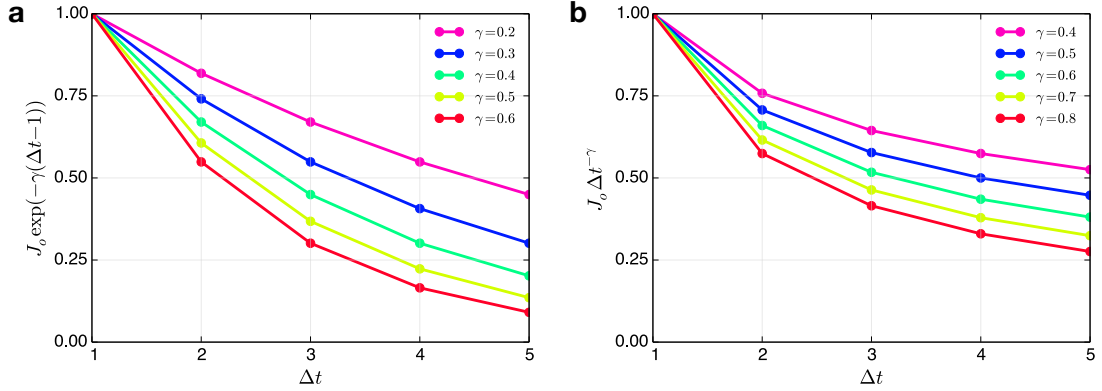


Figure S3: Temporal coupling between two temporal slices as function of the decay parameter γ . By definition the decay is zero between consecutive slices ($\Delta t = 1$). **a**, Exponential decay. **b**, Power law decay.

S3.1.1 Partitioning the dendrogram

The method described above iteratively constructs a dendrogram where temporally localized groups are hierarchically clustered. To extract meaningful social structures we need to partition the dendrogram (Fig. S4). Modularity and partition density have previously been applied for similar purposes [11, 12], but these do not generalize well for temporal processes.

Instead we consider the cluster stability with respect to local and global measures. Local stability (η) is calculated from the average node-wise overlap between consecutive slices (Fig. S5a), while global (σ) is calculated from the average overlap between all slices and the aggregated structure (Fig. S5b):

$$\eta = \frac{\sum_{t=t_{\text{birth}}}^{t_{\text{death}}-1} J(g_t, g_{t+1})}{t_{\text{death}} - t_{\text{birth}} - 1}, \quad (\text{S2})$$

$$\sigma = \frac{\sum_{t=t_{\text{birth}}}^{t_{\text{death}}} J(g_t, G)}{t_{\text{death}} - t_{\text{birth}}}, \quad (\text{S3})$$

where t_{birth} and t_{death} are respectively the birth and death of the gathering, g_t is a temporal slice, $G = g_{\text{birth}} \cup g_{\text{birth}+1} \cup \dots \cup g_{\text{death}}$ is the aggregated structure, and J is the overlap ($J = |i \cap j| / |i \cup j|$), defined as zero if the gathering has only existed for one time bin. Palla *et al.* [13] applied a related measure to estimate the stationarity of communities. Varying the partition threshold (Fig. S5c) we observe a maximum in both measures, indicating a regime where gatherings are both temporally and globally stable. Threshold values of $d \geq 1/2$ are, however, problematic since they merge gatherings that split into two equally sized parts together with both parts, or vice versa. In this scenario, we find that the desirable behavior is to declare the old gathering as ‘dead’ and identify two new gatherings as born. Therefore, to achieve optimal stability and to avoid issues with unwanted merging we partition the dendrogram at $d = 0.49$.

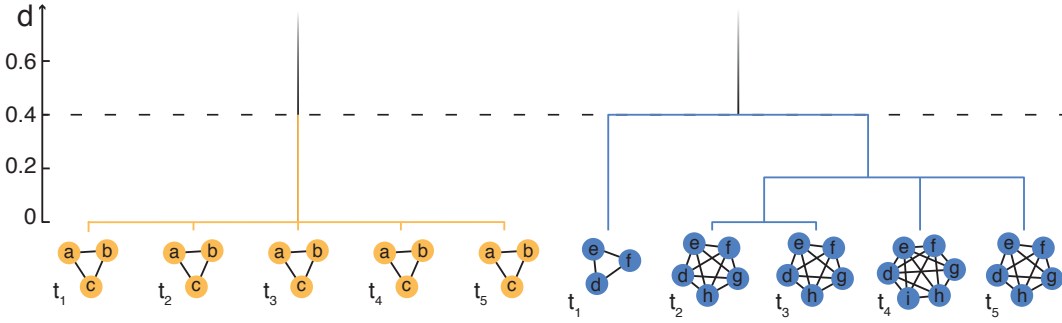


Figure S4: Illustration of constructed dendrogram, depicting distance (d) between groups identified across 5 timebins. The tree is constructed using an exponential decay function with $\gamma = 0.4$. Two gatherings, orange and blue, are inferred by thresholding the tree, where all groups below or equal to the threshold ($d = 0.4$) are merged.

S3.1.2 Temporal decay function

Here we investigate how robust the inferred gatherings are to perturbation of the γ -parameter and using an alternate decay form. To compare two set of gatherings (identified using different γ -values) we calculate the average maximal overlap between individual gatherings

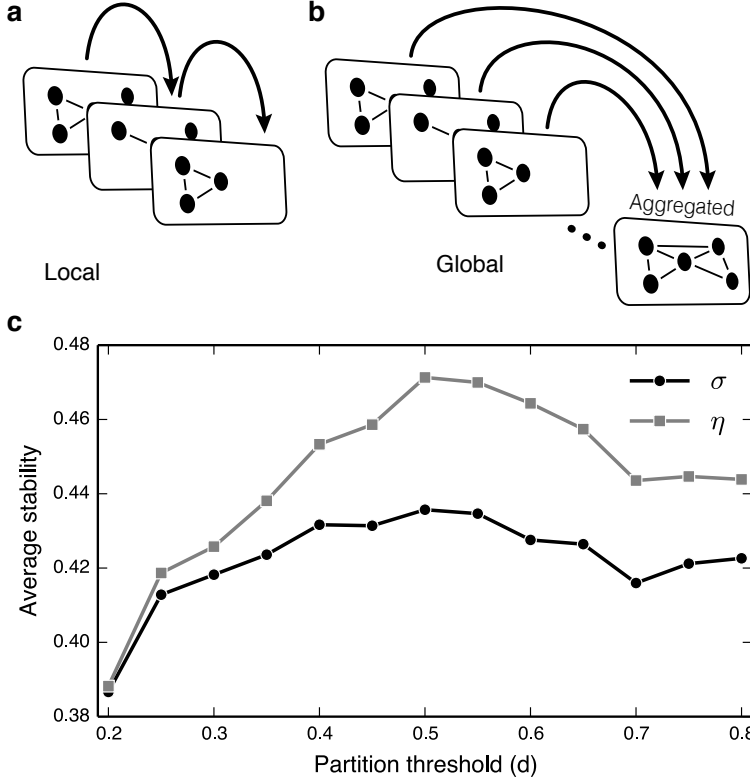


Figure S5: Stability measures of gatherings. **a**, Illustration of the local stability measure, calculated between consecutive slices. **b**, Global measure calculated between each slice and the aggregated structure. **c**, Global (σ) and local (η) stationarity of gatherings as a function of partition value d , averaged over all gatherings identified in January 2014. Gatherings achieve optimal stability around $d \sim 1/2$.

as

$$O_{\gamma\gamma'} = \frac{1}{|G_\gamma|} \sum_{i \in G_\gamma} \max_{j \in G_{\gamma'}} (J(i, j)), \quad (\text{S4})$$

where G_γ denotes the set of gatherings found using a specific value of γ , and $|G_\gamma|$ is the number of gatherings. Overlap is calculated using Jaccard similarity, $J = |i \cap j| / |i \cup j|$, where i and j are respectively gatherings from G_γ and $G_{\gamma'}$. Figure S6 shows the overlap matrix between identified sets of gatherings; in general it assumes overlap values above 0.76 for any choice of parameters. Indicating that the gatherings are robust to even large perturbations.

S3.1.3 Gathering timescales

The outlined method identifies multiple gatherings, some only exist momentarily while others are sustained for long time periods. One can easily imagine brief encounters between good friends as being more meaningful than prolonged interactions between individuals commuting to work; this therefore raises the question of which meetings are meaningful and which are not.

Here we simply adopt the convention developed by the *Rochester Interaction Record* [14], where meaningful encounters are defined as those lasting 10 minutes or longer. In order to filter out spurious connections, we impose the requirement that a gathering must be observed for at least 4 consecutive time slices to be represented in our statistics.

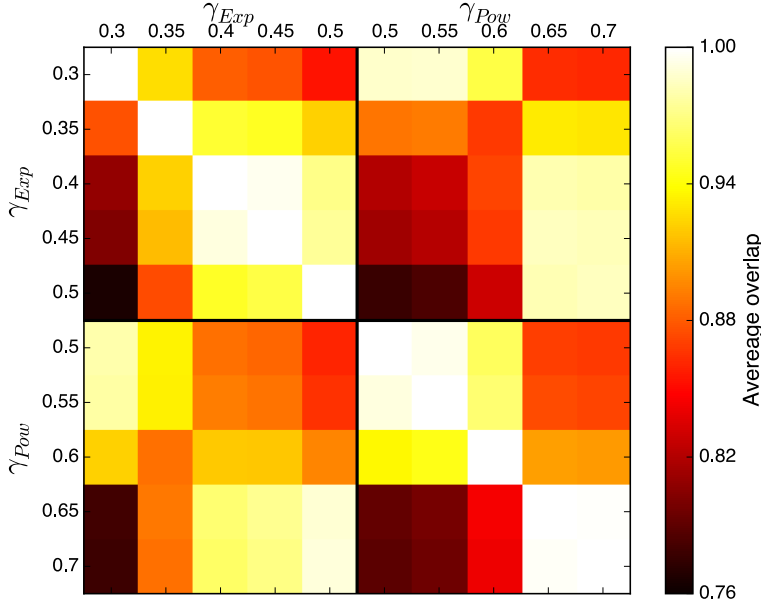


Figure S6: Effect of decay parameter on gathering robustness, calculated using Equation S4. Color-bar shows average overlap between two sets of gatherings, and never drops below 0.76, indicating the robustness of the procedure with respect to parameter perturbations.

While dynamics on 20 minutes+ timescales describe the overall evolution of a gathering, micro dynamics on 5-minute scales represent everyday events such as going to the bathroom. Gatherings, therefore, might disappear and reappear within very short time-intervals (Fig. S7a), in such cases we use imputation, see Fig. S7b.

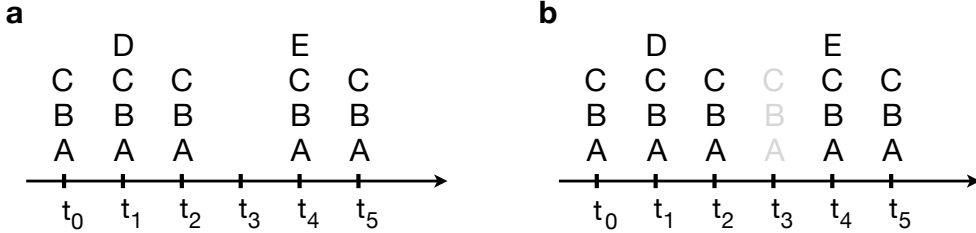


Figure S7: Gathering micro fluctuations. **a**, A gathering appears, disappears, and re-appears moments later. **b**, If this occurs in rapid succession we treat the gathering as if it was present in bin t_3 with the same nodes as in bin t_2 .

S3.2 On- & off-campus gatherings

Gatherings are not geographically constrained and therefore free to occur anywhere, in this section we focus on distinguishing between gatherings that occur on- and off-campus. We do this by applying data acquired through the WiFi channel, where each mobile phone scans for nearby wireless network access points (AP) every 5 minutes or less and logs their unique identifier and name. The entire university campus is densely covered by wireless networks and it is therefore highly unlikely that students located on campus will not see an university AP. Because the names of campus APs follow a uncommon and standardized

naming scheme we can infer when a student is close to a campus access point and use this as a proxy of being on or off campus. For each participant we construct a 5-minute binned vector containing binary values 0 (not on campus) and 1 (on campus). Because gatherings are an ensembles of people we perform majority voting across nodes for each time-bin to determine whether the gathering was on campus or not in that specific time-bin. Since gatherings are spatio-temporal entities, we also perform a majority voting across all time-bins to achieve a hard split and determine whether a gathering mainly occurred on- or off-campus. This yields 13 872 off- and 9 195 on-campus gatherings and with an 92.84% average agreement between votes. The primary reason for disagreement are mobile gatherings traveling to or from campus (e.g. on a bus or walking).

S3.3 Gathering statistics

Gatherings show a broad distribution in both size and duration, see Fig. S8. Dividing gatherings into on- and off-campus categories reveals that meetings occurring on university campus are larger (Fig. S8c), but have considerably lower probability of lasting longer than 4 hours (Fig. S8d). This suggests that large meetings are mainly driven by the class schedule, while meeting duration is determined by social context. Because meetings occurring outside of campus often require increased coordination, our hypothesis is that once groups meet, they will interact across longer periods of time for the meeting to pay off with respect to the organizational cost.

It is also interesting to consider the duration of each meeting as function of the total number of nodes that participate in it. Figure S9a shows broad distributions of duration across all sizes, however, both the mean and median are quite stable and reveal that small gatherings on average have shorter durations compared to larger meetings. Further dividing the data into on- and off-campus categories (Fig. S9b) shows that small meetings on and off campus are quite similar with respect to duration, while larger meetings tend to last longer, provided that they occur off-campus.

Combining the statistics above with Eqs. (S2-S3) we can explore how gathering stability depends on size and duration of meetings. Figure S10a reveals that the local churn between consecutive time slices is quite constant, irrespective of size, indicating that there is a low turnover of nodes between slices—on average much lower than predefined by the partition threshold. For small gatherings, however, we observe finite size effects due to the partition threshold (see sec. S3.1.1). Global stability is lower, but also fairly independent of gathering size. With respect to duration (Fig. S10b), local stability increases as meetings duration increases, revealing that longer meeting have lower turnover of nodes between consecutive slices. The global measure shows similar behavior. It achieves a slightly lower stability and shows that gatherings are globally stable independent of duration. This combination of a constant turnover between slices and a global stability suggests that gatherings contain groups of highly interacting individuals, that are present throughout the entirety of the meeting, while other individuals participate infrequently and are constantly being replaced. Similar social structures have previously been observed in the social science literature where the individuals have been defined as core members [15].

Finally we can look into specific temporal patterns with regards to when on and off campus gatherings occur, i.e. in which 5 minute time-bins they first appear and later

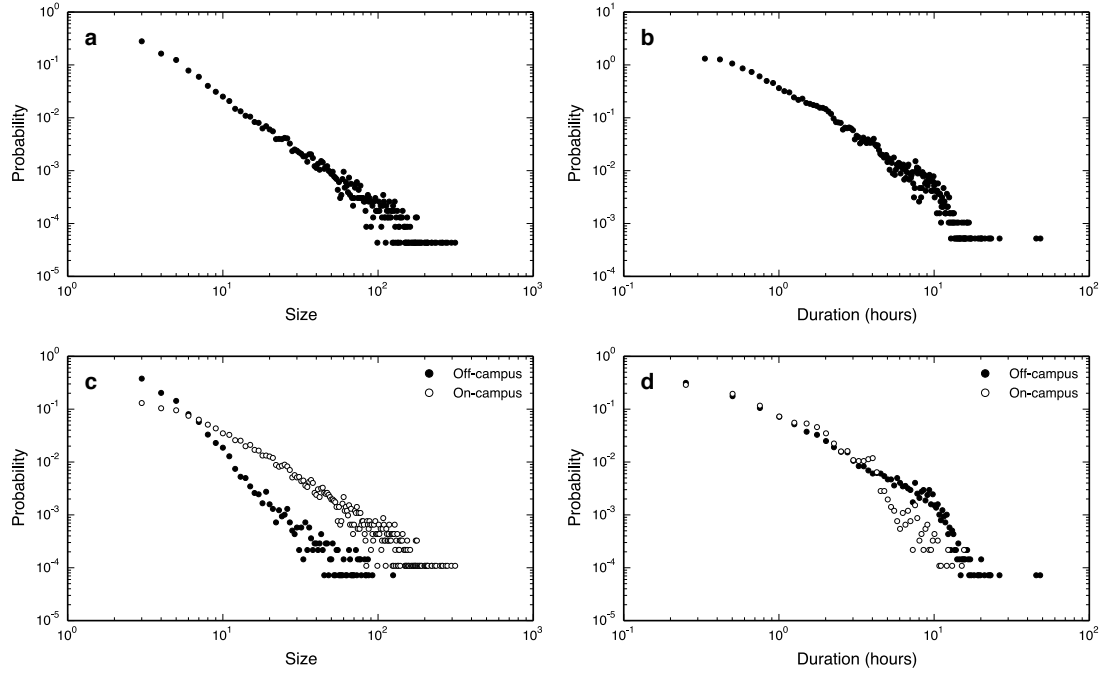


Figure S8: Statistical features of gatherings. Summarizing size and duration distributions for all gatherings (panels a-b) and conditioned on off/on-campus meetings (c-d). **a**, Gathering size distribution. **b**, Gathering duration distribution. **c**, Gathering size distribution for on- and off-campus meetings. **d**, Gathering duration distribution, based on location (on/off-campus).

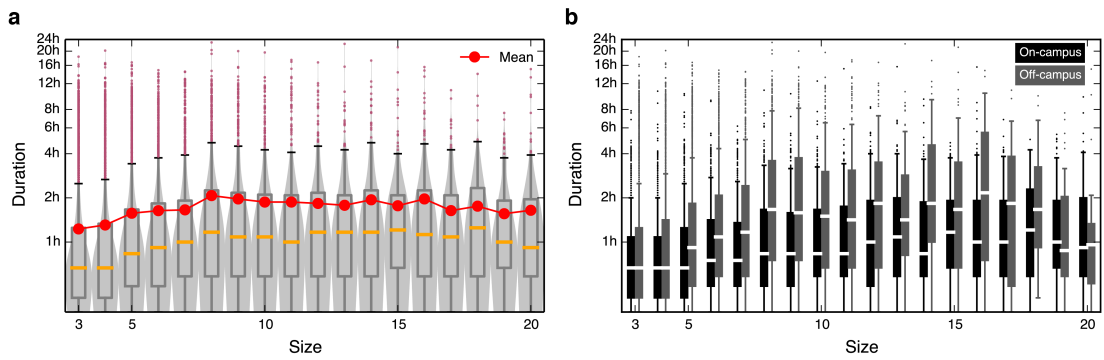


Figure S9: Duration as function of gathering size. **a**, Violin plot shows the distribution of durations as function of size, summarized across all gatherings. **b**, Box plot of the duration distributions divided into on- and off-campus meetings.

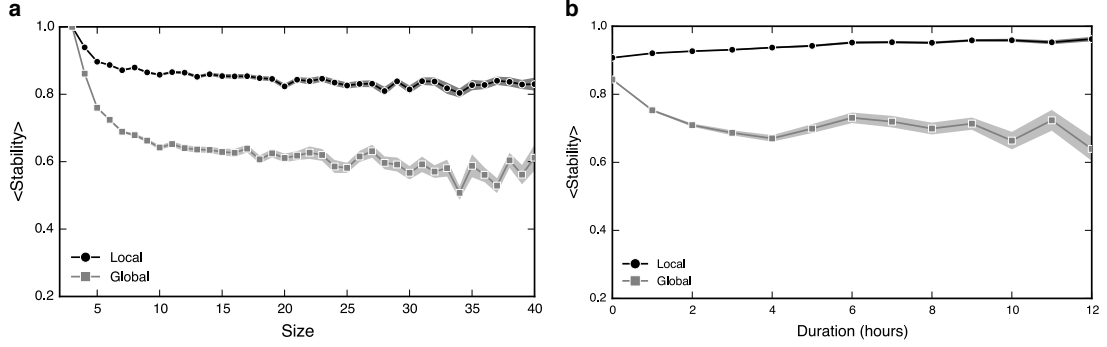


Figure S10: Global and local gathering stability. **a**, Average stability as function of size, averaged across all gatherings with a specific size. Full lines denotes mean, while shaded areas shows the standard deviation of the mean. **b**, Average stability as function of duration, binned into one hour wide bins. Fully drawn lines denote the mean, while shaded areas illustrates its deviation.

disappear. Figure S11a reveals that on-campus gatherings have increased probability of occurring exactly on the hour, while off-campus meetings are evenly distributed—clearly showing the effect of the class schedule. A similar case is seen for the probability of dissolving (Fig. S11b), where on-campus meetings mainly end on integer hour values.

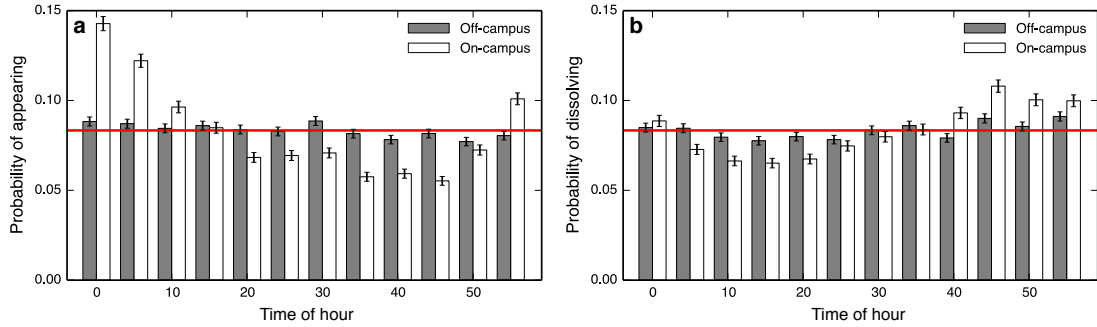


Figure S11: Summary statistics of gathering temporal patterns, summarized for on- and off-campus meetings. Red line denotes the uniform probability distribution for the case where all states are equally probable. **a**, Probability of a gathering appearing, calculated from the first time-bin in which we observe it. **b**, Probability of dissolving, i.e. the last time-bin where we observe a gathering.

S3.4 Temporal communities

So far each gathering only contains information about its local appearance, to gain a dynamical picture we match gatherings across time. Due to soft boundaries, a strict matching criteria is not a feasible method, since a person who coincidentally walks past a group might be included in it. Thus, we expect noise to be present in each gathering. To mitigate this effect we instead match gatherings according to the participation levels of their constituent nodes. Counting the fraction of times a nodes has been present in the gathering

we construct a normalized participation profile, see Fig. S12a. Again, gatherings are matched according to their individual participation profiles using agglomerative hierarchical clustering. Since nodes, however, no longer assume binary values but may assume participation levels in the interval $0 < n_i \leq 1$, we calculate distance based on a continuous version of the Jaccard similarity:

$$D(G_i, G_j) = 1 - \frac{\sum_{n=1}^N \min(G_i, G_j)}{\sum_{n=1}^N \max(G_i, G_j)}, \quad (\text{S5})$$

where G_i is a vector containing node-wise participation values for gathering i , and N is the total number of nodes in $G_i \cup G_j$. The two functions \max and \min act piecewise on the two vectors, and $D(G_i, G_i)$ is defined as 1 between two gatherings that have zero overlap. When merging clusters of gatherings (\mathcal{G}) we apply the average linkage criterion and define the average distance between them as

$$D(\mathcal{G}, \mathcal{G}') = \frac{1}{|\mathcal{G}||\mathcal{G}'|} \sum_{G \in \mathcal{G}} \sum_{G' \in \mathcal{G}'} D(G, G'), \quad (\text{S6})$$

where $|\mathcal{G}|$ denotes the cardinality of a set of gatherings. Other linkage criteria can also be used, such as complete or Ward-linkage. Iteratively this method builds a dendrogram with gatherings as leafs. Thresholding the tree partitions similar gatherings together into communities. A community then consists of all nodes from its constituent gatherings, but it also inherits their individual participation profiles. Thus we need a method to construct a community participation profile from its subcomponents. This can be done using two methods: weighted or unweighted, with the differences illustrated in Fig. S12b. The unweighted method takes into account the gathering participation profiles and calculates the average, weighing each gathering equally:

$$C = \frac{1}{|\mathcal{G}|} \sum_{G \in \mathcal{G}} G. \quad (\text{S7})$$

The weighted version instead assigns each gathering a weight according to its lifetime (τ_{life}), i.e. number of temporal bins it has been present:

$$C^{\text{weighted}} = \frac{1}{\sum \tau_{\text{life}, G}} \sum_{G \in \mathcal{G}} \tau_{\text{life}, G} G. \quad (\text{S8})$$

Both measures comparatively construct similar dynamical communities and yield similar overall statistics; we choose the the weighted version, because it is slightly less influenced by noise.

S3.4.1 Optimal clustering partition

Applying hierarchical clustering to merge similar gatherings into communities again leaves one open question: Which partition value is the optimal? Using a similar line of reasoning to what we presented in in Sec. S3.1.1, we can argue that a threshold value of $D = 0.49$, is

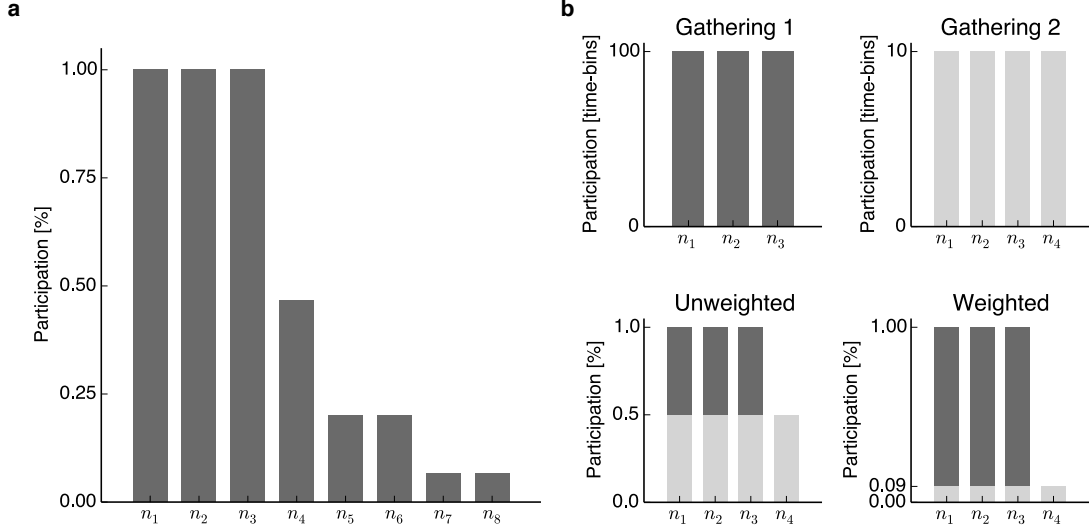


Figure S12: Illustration of gathering participation profiles. **a**, Each vertical bar indicates the overall fraction of time a node has spent in the gathering. Values are normalized according to the total gathering lifetime. **b**, Illustration of how to construct a community participation profile based on its constituent gatherings. Top figures depict two gatherings with unequal lifetimes. Bottom plots illustrate the principles behind the unweighted and weighted methods. Bar colors indicate the role each node plays in the community profile.

preferable. It is, however, possible to estimate the exact optimal threshold value, which we can compare to our hypothesized guess. Applying a heuristic inspired by the Gap statistic introduced by Tibshirani *et al.* [16], we compare the clustering according to a null model distribution (for a comprehensive survey of methods see Milligan and Cooper [17]). Given a total of m gatherings clustered into k clusters (communities): C_1, C_2, \dots, C_k we calculate the within-dispersion measure as,

$$W_k = \sum_{r=1}^k \frac{1}{2|C_r|} \sum_{i,j \in C_r} D_{ij}, \quad (\text{S9})$$

where D_{ij} is defined in equation S5 and denotes the pairwise distance between gatherings i and j that both belong to cluster C_r . Again $|\cdot|$ denotes the cardinality of a cluster, i.e. the number of gatherings that are clustered in C_r . The factor 2 takes double counting into account. Thus W_k is the accumulated within cluster sum of differences around the cluster mean. Applying the principles from Tibshirani *et al.* [16] we compare $\log(W_k)$ to an expected value generated by a null model distribution of the data. The gap measure is defined as

$$\text{Gap}(k) = 1/B \sum_{b=1}^B \log(W_{kb}) - \log(W_k), \quad (\text{S10})$$

where B denotes the number reference data sets. The optimal number of clusters is then the value of k for which $\log(W_k)$ falls furthest below the reference data curve, i.e. the value

of $k = k^*$ such that

$$k^* = \underset{k}{\operatorname{argmin}} \{k | \operatorname{Gap}(k) \geq \operatorname{Gap}(k+1) - \tilde{s}_{k+1}\}, \quad (\text{S11})$$

where $\tilde{s}_k = s_k \sqrt{(1 + 1/B)}$ and s_k is the standard deviation of $\log(W_{kb})$ over the B synthetic datasets. Each null model is constructed by assigning random participation values, chosen from a uniform distribution, to random nodes, thus creating reference gatherings with similar size distributions. According to Fig. S13 the gap statistic achieves a minimum, indicating that the optimal place to cut the dendrogram is at distances of $D = 0.50$, in good agreement with the previously stipulated value. A threshold value of $D = 0.50$ is, however, problematic since it will cluster gatherings with 50% overlap. To avoid this issue we cut our dendrogram at $D = 0.49$, merging 23 067 gatherings into 7 320 distinct dynamic communities.

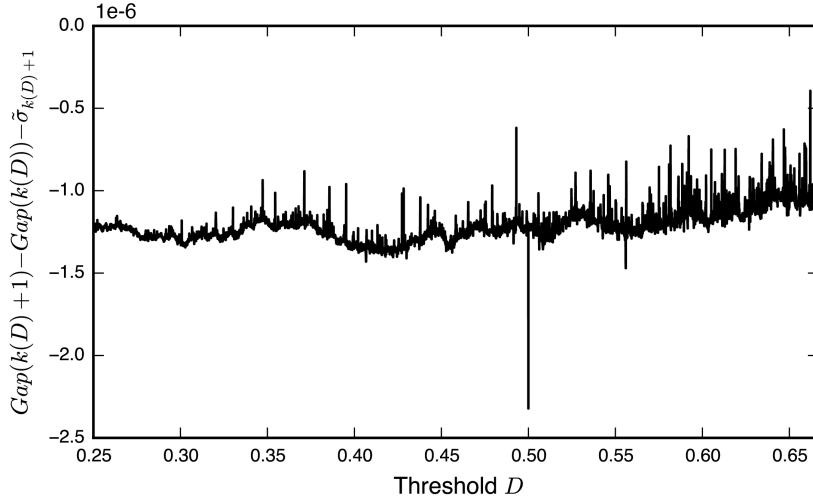


Figure S13: Estimating the optimal number of clusters using the gap measure, with the number of clusters k being directly related to the threshold value D .

S3.5 Dyadic relations

Dyads are a special case of social relations and are hypothesized to be qualitatively different from groups [7, 8], but the way we identify them is similar to the outlined framework in Section S3.1. Since dyads can only be components of size two, tracking their evolution is trivial, because we can apply a strict merge criterion, i.e. require 100% overlap. This also applies in the case when tracking repeated appearances (dyadic cores) across the full duration of the dataset. Thus, the only distinction between a gathering and a dyadic gathering lies in the fact that size of the gathering will always be fixed, meaning if a dyad evolves into a group (of any size larger than two) then we claim that the dyad in question has ended and a new group relation has been initiated. In accordance with gatherings, we require dyadic gatherings to be present for at least four consecutive bins (20 minutes) to be

considered as meaningful. In total we observe 34996 dyadic gatherings and 4844 unique dyads.

S3.5.1 Dyad statistics

Similar to group relations, dyadic gatherings produce a broad distribution of durations comparable to the gatherings of groups (Fig. S14a). In addition, dyads also produce a broad distribution of repeated appearances, see Fig. S14b, with a majority of gatherings only being observed once, while others on average can appear multiple times per day.

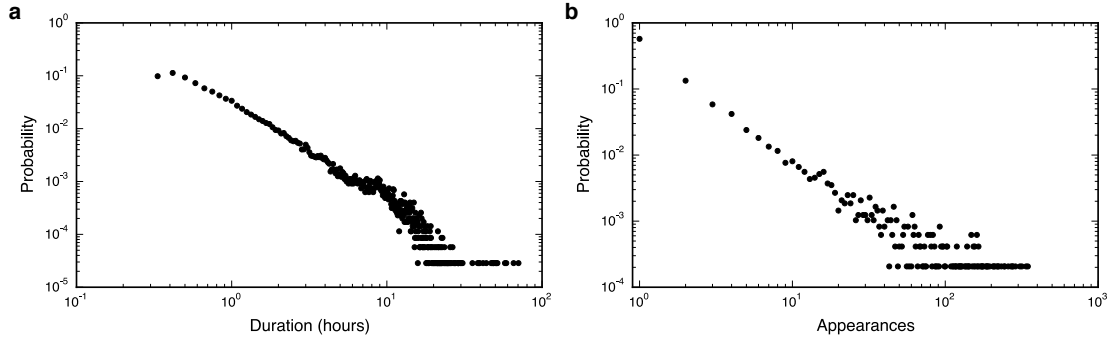


Figure S14: *Statistical features of dyadic gatherings. a, The distribution of duration for dyadic gatherings. b, The distribution of the number of appearances.*

S4 Cores

In the previous section (see Fig. S10b) we found that certain individuals are present for a majority of a gathering’s lifetime. In this section, we begin by describing how such *cores* are extracted from temporal communities. Then, we describe the fundamental statistics of cores (distribution of appearances, distribution of sizes, individual membership in cores, subcore-statistics), as well as the key differences between work- and recreational cores. Finally, we study the temporal patterns and entropy of core-appearances.

S4.1 Extracting cores

Nodes within each community have varying attendance (see Fig. S12a), some are only members for a limited time, while others interact over extended periods of time. Thus, participation profiles show pronounced core structures, highlighting individuals that act as ‘generators’ of each community.

Consider the participation levels of individuals as ordered profiles (Fig. S15a), where each bar denotes the fraction of time a node has spent in a community relative to the community’s total lifetime. A significant gap in this profile identifies core nodes. We compare this to a participation profile generated by a random process (Fig. S15b), where we pick a random participation level between 0 and 1 (for each node) from a uniform distribution. The maximal gap in this random profile thus tells us whether the real gap is

significant. We can estimate the average expected gap size and deviation by generating many ($N = 10\,000$) random participation profiles. Generalizing this notion to all sizes of communities we evaluate how significant a gap is compared to the expected value generated at random. The decision boundary in Fig. S16 divides gap sizes into two regions. If the actual gap is greater than the average null-model gap μ_{random} plus one standard deviation σ_{random} , we define the core to be significant. Thus, we only keep cores with gap sizes above $\mu_{\text{random}} + \sigma_{\text{random}}$. According to this criterion, we find that 7 146 out of the 7 320 (97.6%) inferred communities display a pronounced core structure.

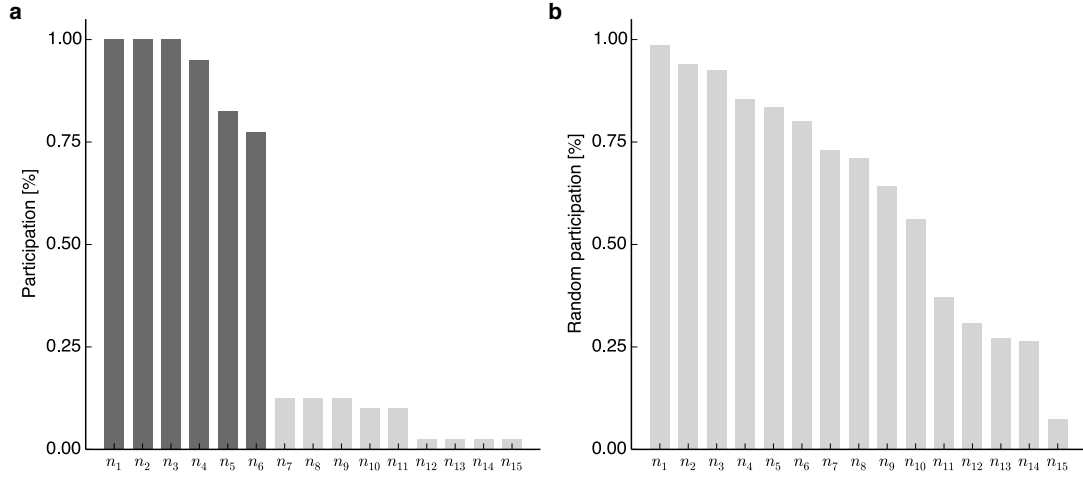


Figure S15: Extracting cores from community participation profiles. Dark gray bars denote nodes with participation levels above the maximal gap. **a**, Ordered participation profile for a community composed of 15 individuals. **b**, Similar as in panel a but generated from a uniform random distribution.

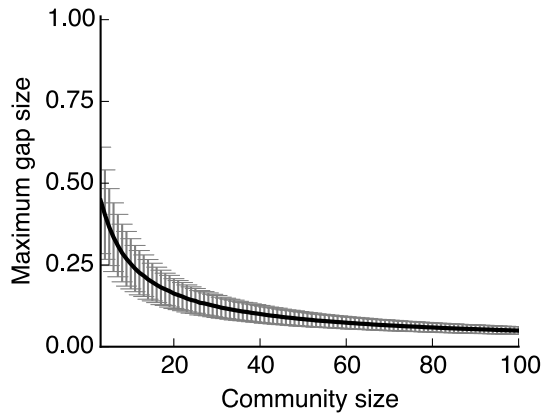


Figure S16: Core selection boundary. Decision boundary calculated from $N = 10\,000$ independent trials for each size. Black line denotes the mean max gap value while error bars indicate the standard deviation.

S4.2 Core statistics

Cores have a broad distribution of the number of appearances (Fig. S17a), ranging from cores appearing only once to on average occurring multiple times per day. The size distribution is also heavy-tailed (Fig. S17b).

To produce meaningful temporal statistics we henceforth focus on cores which on average are observed more than once per month, this limits our focus to individuals that appear in these. Fig. S18 shows the distribution of the number of cores each individual is part of; it is a broad distribution with a majority of individuals partaking in few cores while a small minority of users are extraordinary social and have more than 10 cores.

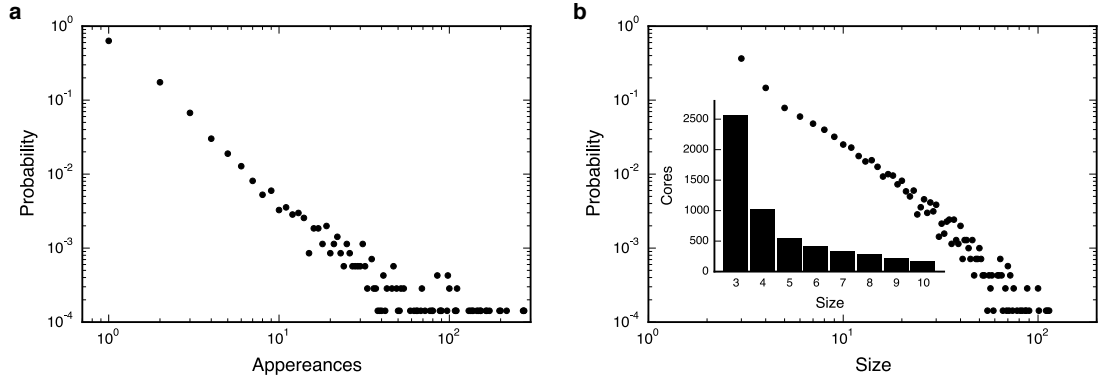


Figure S17: Core statistics, where we focus on cores of size greater than two. **a**, Probability distribution of the number of appearances per core. **b**, Size distribution. Inset shows raw numbers for specific sizes.

S4.3 Subcores

Because cores span a wide range of sizes small cores can appear as subcores embedded within larger ones, see example in Fig. S19a. In fact, our methodology allows for and identifies highly overlapping and hierarchically stacked structures. We define a core to be a subcore if, and only if, it is fully contained in the larger one. Figure S19b shows a broad distribution in the number of subcores that are contained within individual cores, with a majority of cores only containing one subcore, while other can contain more than ten. There is of course a dependence on size, such that bigger cores have larger probability of containing more subcores. Figure S19c quantifies this phenomenon by considering the fraction of subcores that cores of size s contain.

S4.4 Work & recreational cores

In Sec. S3.2 we determined for each gathering whether it occurred on or off campus. For cores this distinction is not possible since cores consist of multiple gatherings that in principle can occur both on or off campus. Instead we count the number of times each core appears on and off campus and perform a majority voting. Cores that have a higher frequency of on-campus meetings are denoted as *work* cores, while we call cores that appear

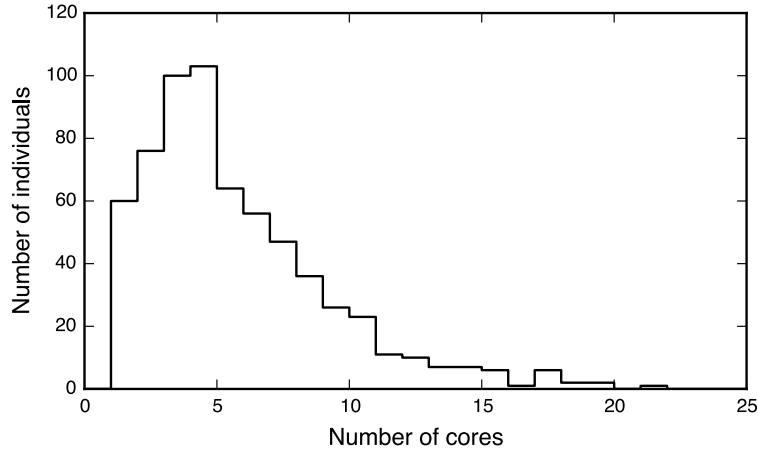


Figure S18: *Distribution of the number of cores per individual. Calculated for users that appear in frequently observed cores, i.e. cores that on average are observed more than once per month. A minority of individuals partake in more than 10 cores, with the distribution being centered around a mean value of ~ 5 .*

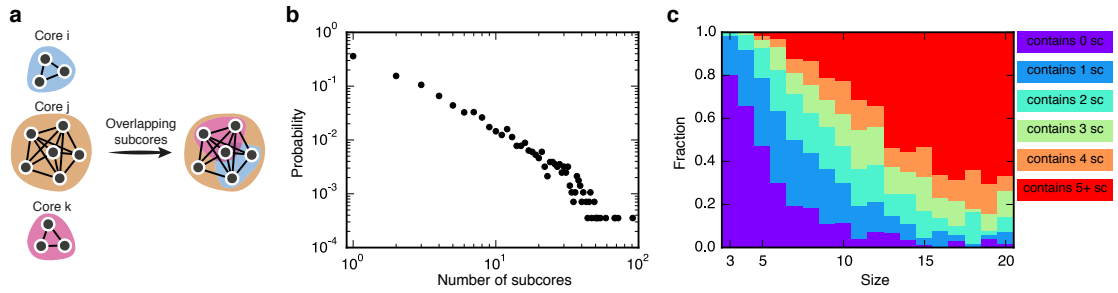


Figure S19: *Subcore structures and statistics. a, Illustration of overlapping and hierarchically stacked core structures. b, The distribution of the number of subcores contained within individual cores. c, Fraction of cores that contain 0, 1, 2, 3, 4, and 5+ subcores (sc) as function of core size.*

more frequently off-campus as *recreational* cores. In case of a tie, we label the core as recreational. Figure S20a shows the voting schedule and indicates the split, while Fig. S20b shows the waiting time probability between consecutive meeting events. For work cores the waiting time probability shows clear signs of daily and weekly patterns, suggesting that these cores may be driven by the class schedule. Recreational cores on the other hand exhibit a more subtle pattern that slowly decays and is considerably higher during nighttimes—suggesting that two fundamentally different mechanisms drive the activity of the two groups. Splitting the number of cores per individual (Fig. S18) up into the two categories yields the results shown in Fig. S20c-d, where users have a broad degree of recreational cores, on average participating in 2.53, while the number of work cores is more localized with an average value of 2.74. According to Fig. S20e-f individuals spend more time in recreational cores than work cores, clearly depicting that context has great influence on the properties of meetings.

S4.5 Meeting regularity

Each core has a specific temporal pattern linked to it denoting the periods of time it has been present, see Fig. S21a for an example. The information contained in each pattern can be quantified using Shannon entropy [18], defined as

$$H = - \sum_t p_t \log_2 p_t, \quad (\text{S12})$$

where the sum runs over all temporal bins t and p_t is the probability of observing a specific core within given time-bin. For each core we aggregate its meeting patterns across the full study duration into weekly 1-hour bins, then within each bin we calculate the probability of observing the core. Entropy is calculated individually for cores using Eq. S12. According to Fig. S21b there is clear differences between the meeting patterns for work and recreational cores. Further, the entropy distributions (Fig. S21c) reveal that recreational cores, on average, have higher entropy and thus lower meeting regularity—indicating that they do not meet within pre-scheduled temporal-bins.

S4.6 Ego viewpoint

So far we have mainly focused on the overall structural and dynamical features of cores, but we can reverse the perspective and look at cores from the perspective of individuals. From this perspective, each core provides a social context for the situation in which a person is embedded, whether it is in a work, recreation, or other relation. Figure S22a shows the involvement of a representative individual that is involved in multiple cores, producing hierarchically nested and overlapping structures. The corresponding temporal pattern of cores (Figure S22b) reveals a complex behavior.

S5 Predicting behavior from routine

In this section we describe the detailed analysis leading up to routine-based geographical location and social context prediction presented in the main text.

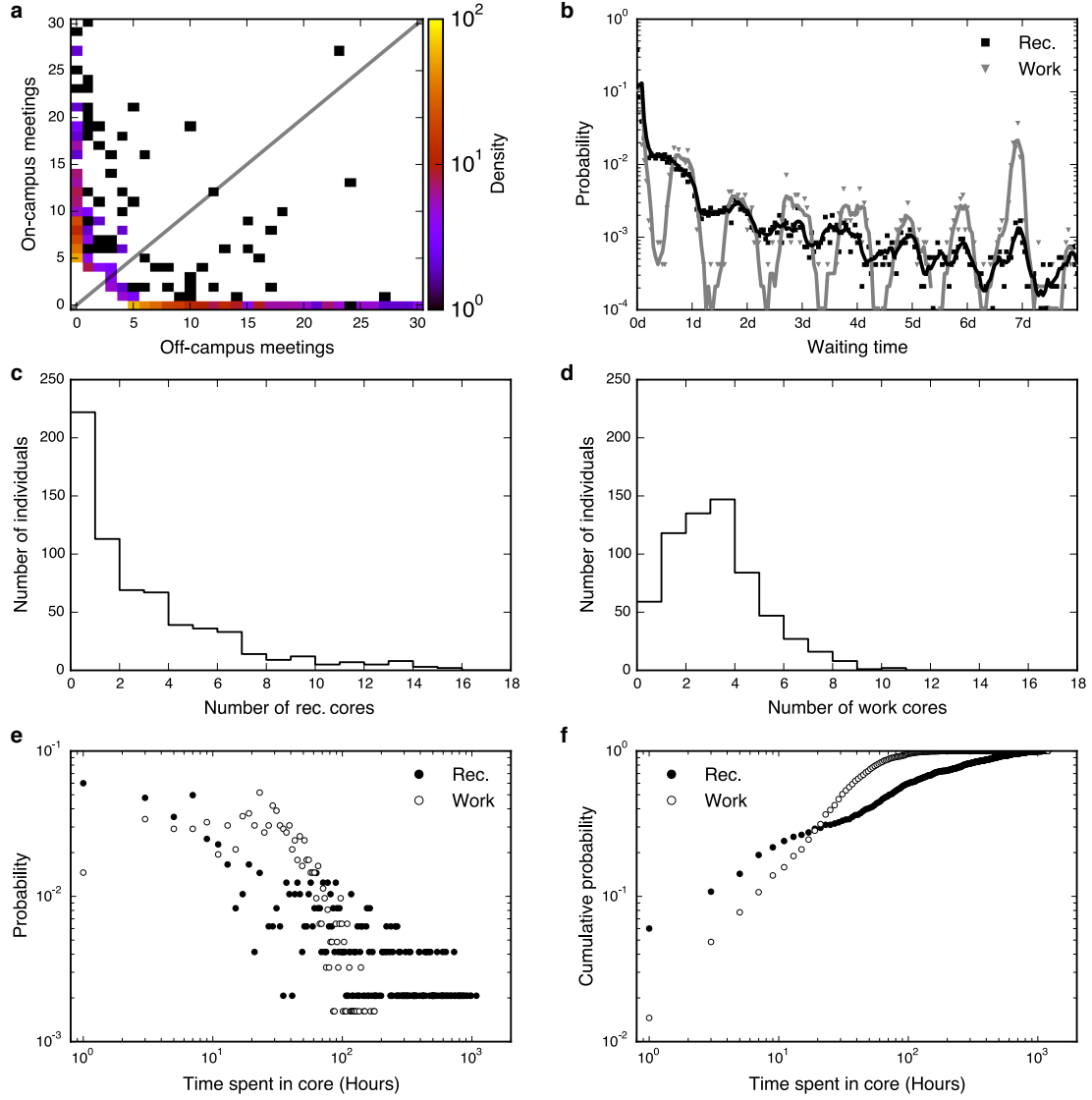


Figure S20: Distinction between work & recreational cores. **a**, Excerpt of voting scheme depicting the number of on- and off-campus meetings for each core. Gray line splits the area into work (above) and recreation (below) categories. **b**, Inter event time distributions between consecutive meetings, aggregated across all cores of similar class (work/recreation). Events are hourly binned and full lines denote moving averages, calculated using 4-hour windows. **c**, Number of work cores per user. **d**, Number of recreational cores per user. **e**, The distributions of how much time individuals, in total, spend in cores with work and recreational context. The average individual spends approximately 38 hours in a work setting, and 120 hours in a recreational context during the period of study. **f**, Cumulative probability distribution of data shown in panel e.

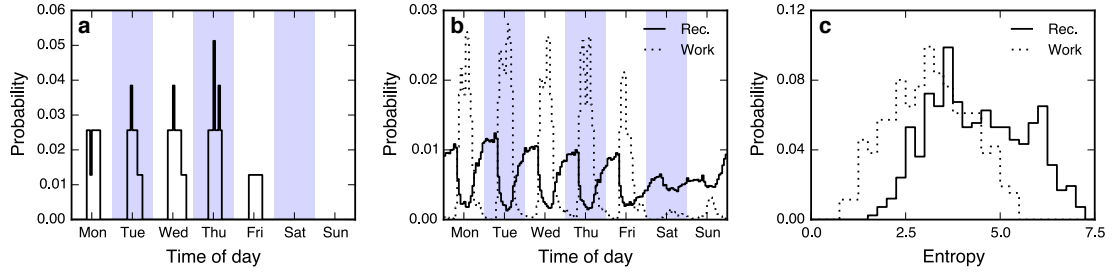


Figure S21: Meeting regularity of cores. **a**, Example pattern for a single core, denoting the probability of observing the core. Data is aggregated across all weeks into 1-hour wide bins. **b**, Aggregated meeting patterns across all cores, showing the probability of observing work and recreational cores. **c**, Distributions of meeting time entropy calculated across all cores and divided up into work and recreation categories.

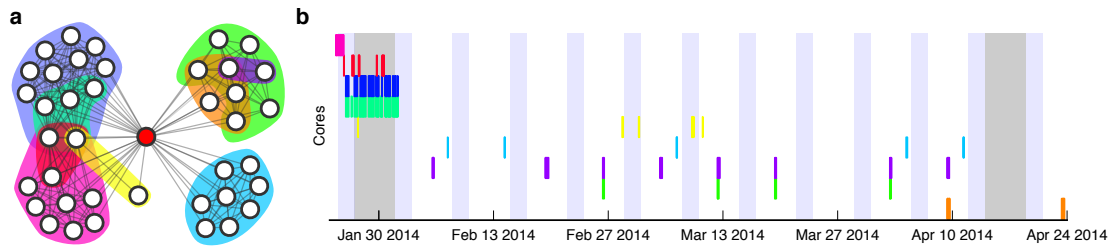


Figure S22: Ego-centric perspective of cores for a representative individual. **a**, Network perspective, revealing overlapping and nested structures. **b**, Temporal dynamics of individual cores, colored accordingly.

Song *et al.* have used entropy, an information theoretic measure in order to estimate the upper bound of the predictability of individuals' mobility patterns [19]. We argue here, that in analogy to geospatial behavior, human social life can be described by a temporal sequence of 'social states'. These can be used to quantify the predictability of social life.

Given a sequence of states for an individual i we can define entropy in two ways. First we can think of predictability in a temporally uncorrelated sense with entropy defined as

$$S_i^{\text{unc}} = - \sum_j^{N_i} p_j \log_2 p_j, \quad (\text{S13})$$

where p_j is the probability of observing state j and N_i is the total number of states observed by person i . Eq. S13 captures the uncertainty of your location history without taking the order of visits into account, thus discarding information contained in the daily, weekly and monthly sequences of behavior. A more sophisticated measure that includes temporal patterns is *temporal entropy*:

$$S_i^{\text{temp}} = - \sum_{T'_i \subset T_i} p(T'_i) \log_2 [p(T'_i)], \quad (\text{S14})$$

where $p(T'_i)$ is the probability of finding a subsequence T'_i in the trajectory T_i . From the entropy one can estimate the upper bound of predictability (Π_i) by applying a limiting case of Fano's inequality [19–21]:

$$S_i = H(\Pi_i) + (1 - \Pi_i) \log_2 (N - 1), \quad (\text{S15})$$

where $H(\Pi_i) = -\Pi_i \log_2(\Pi_i) - (1 - \Pi_i) \log_2(1 - \Pi_i)$.

S5.1 Comparing with previous studies

In the main manuscript (Fig. 3a) we show that our ability to predict the location of individuals has an upper bound of 71% which is significantly lower than the 93% reported by Song et al. [19]. The main reasons behind this discrepancy are discussed below.

S5.1.1 Difference in populations

Our study population is comprised of university students that (1) not necessary have a single home/nightly location, (2) have a rich free time which they can utilize to explore new locations, and (3) have multiple work locations due to classes being distributed across a large university campus.

In contrast, Song et al. studied a sample of 50 000 individuals selected from a total population of ~ 10 million anonymous mobile phone users. The users were chosen according to (1) their travel patterns, where individuals had to visit more than two cell tower locations during the observational period of three months and (2) their average call frequency which had to be $\geq 0.5 \text{ hour}^{-1}$, effectively selecting individuals with at minimum of 12 calls per day.

S5.1.2 Geospatial resolution

Previous studies have applied call detail records (CDR) as proxy for location, inferring the position of individuals depending on which cell tower their mobile phone is connected to during a call [19, 22, 23]. While the granularity of cell tower locations in cities is around 800 meters it can be on the order of kilometers in more rural areas [24]. Figure S23 illustrates the effect of using cell tower data for positioning, as it can cluster otherwise distinct places together as one. Our location data on the other hand has a typical accuracy below 60 meters [1], enabling a more accurate spatial estimation.

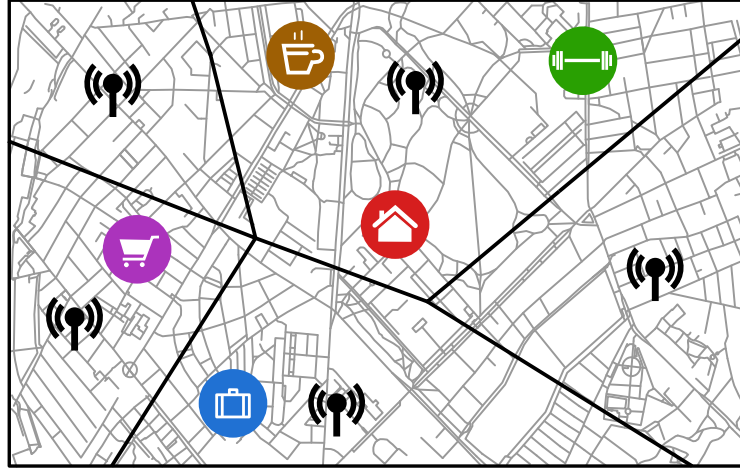


Figure S23: *Cell tower resolution of a city. Otherwise distinct places can be grouped together under the same cell tower, coarse graining the geospatial position and increasing the probability of guessing correct.*

S5.1.3 Effects of binning

Song et al. [19] chose, because of data granularity, to segment their data into one hour time-windows. Since our data has finer granularity, down to the minute scale, we can choose a different resolution, but which is optimal?

We show in Fig. S24a that the finer we segment time, the better we are able to predict your location in the next time-bin. By reducing temporal bin size it is possible to achieve arbitrarily high levels predictability because segmenting data into finer time-windows increases the number of bins, which in turn leads to respective states obtaining a higher frequency of visits.

The bin-size affects both temporal and uncorrelated entropy, however, so one hypothesis is that it is still meaningful to consider the ratio between the temporally uncorrelated entropy and the temporally correlated entropy. To investigate, we study on the ratio S_i^{temp}/S_i^{unc} , evaluated across all individuals i . As is clear from Fig. S24b the ratio shifts towards lower values for small windows, revealing that the choice of bin width has a greater impact on temporal entropy. This suggests that predictability is greatly influenced by the choice of time-window, as has also been noted by [21, 25]. In short, considering the ratio

S_i^{temp}/S_i^{unc} does not solve the binning problem. Ultimately this implies that the smaller bins we apply the better we are at predicting. Because we currently are unaware of any timescale that is fundamentally descriptive of human behavior, we chose to work with temporal sequences in their natural form instead of segmenting them, predicting ‘next state’ rather than ‘state of next time bin’ (see Fig. S25).

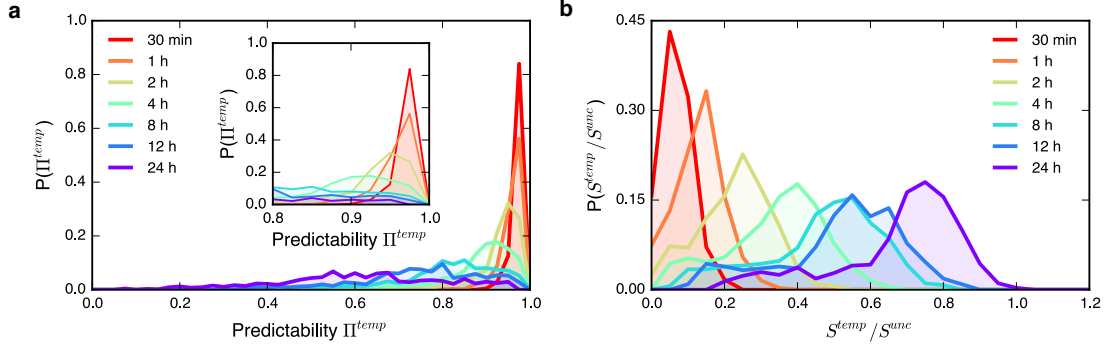


Figure S24: Effects of binning on predictability bounds. **a**, Predictability as function of window size. Inset shows a close up for high values of predictability. Segmenting time into finer bins yields higher bounds of prediction. **b**, Ratio between temporal and uncorrelated entropy values. As the size of time-windows is narrowed the ratio shifts towards zero, indicating that $S^{temp} \ll S^{unc}$.

S5.2 Data for prediction

S5.2.1 Social vocabulary

We describe the social life of an individual based on the context provided by cores, where we focus on individuals that appear in frequently observed cores (observed, on average, at least once per month). In order to include the full social life of an individual we incorporate the context provided by dyads as well as the the information contained in infrequently observed cores. Note that if a core/dyad is only observed once then we denote it as a ‘noise’ state. In addition we construct one supplementary context: ‘alone’ denoting periods of time where an individual is not socially active. Following Song et al. [19] we construct a time-series of social contexts for each user. However, as noted above, we do not segment the sequence into temporal bins, but keep it in its natural form, and predict ‘next state’ (illustrated in Fig. S25) for both geospatial and social prediction.

S5.2.2 Location vocabulary

In addition to social context, we also collect geographical traces for each user, enabling us to reconstruct their mobility patterns. To infer context from raw location traces we use the same definitions as Cuttone *et al.* [26], where a point of interest (POI) is a location of relevance for a person, such as home, work, or a cinema. POIs are inferred by applying a density based clustering algorithm [27], with a density grouping distance of 60 meters and requiring stops to consist of at least two samples, meaning that a person must have spent a minimum of 15 minutes in the same location.

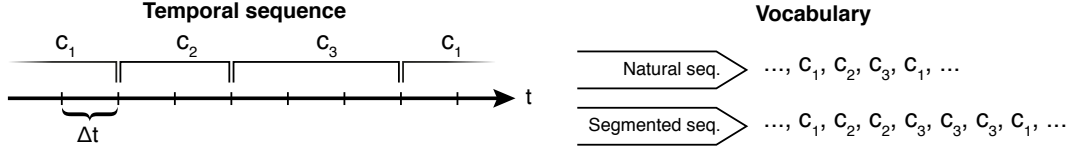


Figure S25: Time-series of states. Left part of the figure presents social states as they naturally occur in a temporal sequence, with Δt denoting a segmentation of the sequence within arbitrary sized time-windows. Right panel illustrates the difference in vocabularies, where the ‘natural’ sequence focuses on the order of states, while ‘segmented’ also weights states according to their duration.

S5.2.3 Convergence of states

Figure S26 shows the distribution of the number of distinct social and geospatial states for increasing windows of time. After 90 days both probability distributions converge, implying that the number of states visited by users is saturated, indicating that we can uncover a majority of states frequented by individuals. We, however, expect this saturation only to be meta-stable, because the social networks change across adulthood [28].

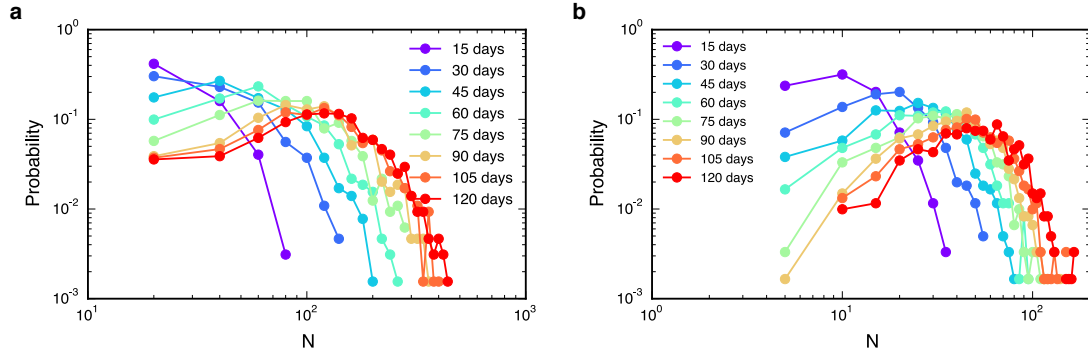


Figure S26: Distribution of the number of distinct states within a time window. **a**, Convergence of the number of social social states, showing saturation after a time window of 90 days. **b**, Saturation of visited locations, also convergence after 90 days.

S5.3 Prediction

Following Eq. S13-S15 we first calculate the respective entropies of the behavioral patterns for each individual. We show in Fig. S27a that our social patterns have lower entropy than our mobility. The figure shows that an average person approximately occupies $2^2 \approx 4$ social states and $2.5^2 \approx 6.25$ location distinct states. We also find that humans are potentially more predictable based on their social contexts than their past locations, see Fig. S27b. Previous studies have found higher levels of predictability [19, 29], see Sec. S5.1 for a full discussion.

Fig. S28a shows the interrelation between S_{social}^{unc} and $S_{location}^{unc}$, surprisingly there is no correlation between the two measures (Spearman correlation $\rho = 0.053$, p -value = 0.191), indicating that humans can be highly predictable in a social sense but very unpredictable

location-wise and vice versa. A similar lack of correlation is observed for S^{temp} (Spearman correlation $\rho = -0.008$, p -value = 0.84), see Fig. S28b.

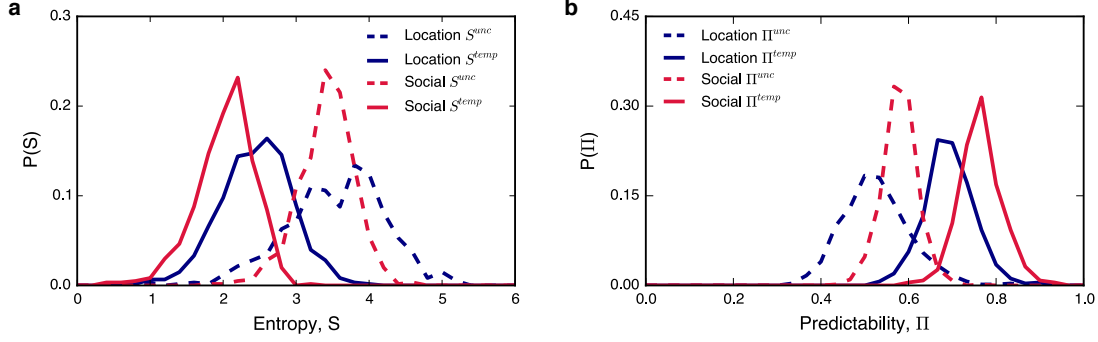


Figure S27: Probability distributions of entropy and predictability. **a**, Distributions of temporal and uncorrelated entropy for location and social behavior. As expected, temporal patterns contain more information than just frequency of visits, and hence have a lower entropy. **b**, Predictability distributions for uncorrelated and temporal patterns. On average, our social behavior is more predictable than our geospatial behavior.

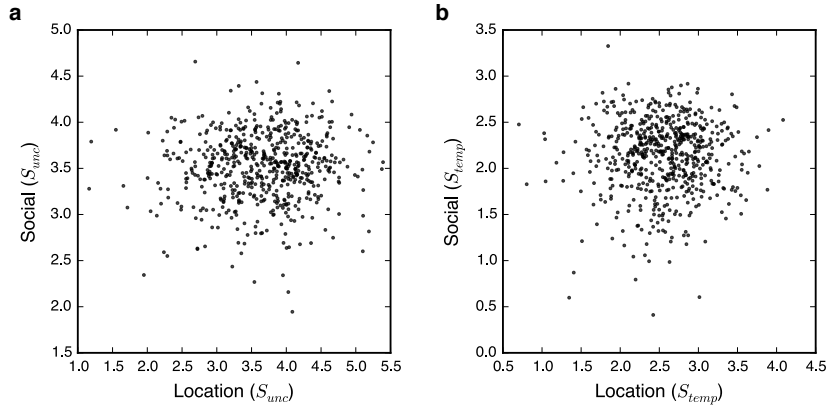


Figure S28: Correlation between social and location entropy. **a**, Mutual dependence between uncorrelated entropy for social and location states per individual. **b**, Correlation between temporal entropies for social and location states.

S5.4 Temporal aspect of predictability

According to Fig. S28 there is no correlation between social and geospatial aspects of human life, but this is measured in terms of overall dependence. In real life we have varying degrees of predictability, a simple way of visualizing this is to look at the number of states as function of time. According to Fig. S29a we have a low number of location states during nights (resulting in low entropy) because we mainly sleep at a single location, while during days and evenings our entropy is higher, in part because we occupy more states. Note that there is very little overlap between locations visited in each 8-hour bin, e.g.

morning locations are different from day locations. On Fridays we visit more locations than any other day, while during weekends we are more stationary. If we, however, consider the total number of distinct visited places (Fig. S29b) we see that Fridays and Saturdays are special because those days are used to explore new locations. Therefore, predicting location during weekends based on routine is more difficult, since we have higher entropy during these periods. Our social behavior (Fig. S29c) resembles our mobility, where we socialize mainly during the day and less during the night. Weekends are again special, interestingly we here observe a drop in the number of social states, because we are not required to go to work or school. Fig. S29d shows that the number of social states decreases during weekends, meaning our participants reserve weekends to socialize with a few selected friends.

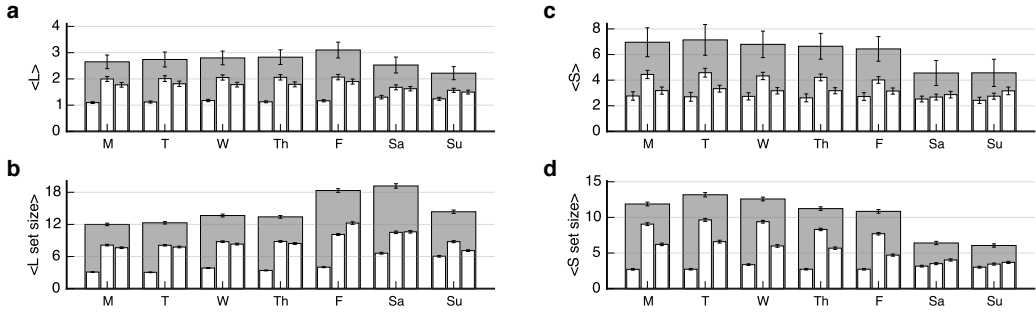


Figure S29: Nested histograms showing the temporal aspects of predictability. Binned using daily and 8-hour intervals (12 am - 8 am, 8 am - 4 pm, 4 pm - 12 am), outer bars (gray) denote days while inner bars (white) denote 8-hour windows. Bars do not necessarily add up, because one can have an overlap of states between the 8-hour bins. All values are averaged across the student population. **a**, Number of average observed locations per bin. **b**, Total number of visited distinct locations. **c**, Average number of social states per time-bin. **d**, Total number of distinct social states.

We quantify the relation between social and geospatial traces by looking at the normalized mutual information. Mutual information is a measure of the variables' mutual dependence, i.e. how much knowledge of one variable reduces uncertainty about the other. It is defined as

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right), \quad (\text{S16})$$

where $p(x)$ is the probability of observing an individual in state x . It is symmetric $I(X, Y) = I(Y, X)$ and nonnegative $I \geq 0$, and mutual information is zero if and only if $p(x)$ and $p(y)$ are independent such that $p(x, y) = p(x)p(y)$. Normalized mutual information is defined as [30, 31]:

$$I_{\text{norm}} = \frac{2I(X, Y)}{H(X)H(Y)}, \quad (\text{S17})$$

where $H(X) = -\sum_{x \in X} p(x) \log(p(x))$ is the uncorrelated entropy of a behavioral pattern. According to Figure S30a our social and geospatial behaviors are correlated during the week, where we tend to meet the same people in the same places. During weekends this correlation between our social and location behavior is lower. Thus, while the geospatial behavior during weekends becomes more unpredictable (Fig. S30b) our social patterns

become simpler (Fig. S30c); consistent with Fig. S29 and previous work [32]. In our population, periods of geospatial exploration are associated with social consolidation, a non-trivial phenomenon which we exploit in the next section.

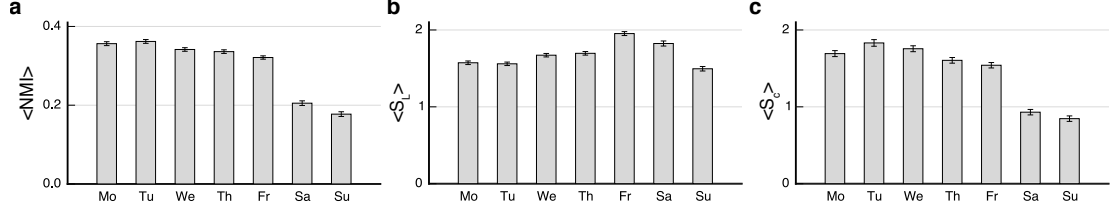


Figure S30: Relation between social and geolocation traces. **a**, Average normalized mutual information per day, clearly depicting a difference between weekdays and weekends. **b**, Average entropy for geolocation patterns. **c**, Average entropy for social patterns.

S6 Social prediction

It has previously been shown that spatial behavior between individuals that share a social tie is correlated [33–36]. But the onset of co-presence lacks a temporal signature, so while the spatial traces overlap it is not generally possible to specify exactly when a friend is predictive for an individual’s behavior. Cores, however, do provide such context. An incomplete set of members provides a clue that a social interaction is about to occur (i.e. the final group member is about to arrive).

We test this concept on cores of size three; thus provided we observe two members we measure the probability of the last member arriving within the next hour. To avoid testing the hypothesis on scheduled meetings we focus on weekday nights (6 pm - 8 am) and weekends. This is the period where routine driven prediction is at its weakest. Further, in order to avoid circularity, we evaluate the hypothesis on a test month (May 2014), during which we have not identified gatherings.

S6.1 Null models

For evaluation purposes we compare cores to two reference models, both generated from real world data. We segment interactions into undirected and unweighted daily graphs, see Fig. S31a. The first null model, which we denote *random*, constructs reference groups by randomly drawing nodes from each daily graph. The second model utilizes a breadth-first search in the daily graph to create reference groups starting from a randomly chosen seed node by searching its local neighborhood. A new seed node is chosen if the search is restricted to components with fewer than three members. *BFS* is a strict null model and requires all nodes to have shared a physical interaction. We disregard reference groups if they happen to be identical to a core.

S6.2 Comparison

We test the hypothesis on a sample of 340 frequently occurring cores and 10 000 reference groups for $n = 100$ independent trials. For 33 cores we never observe an incomplete set of members during the month of May, therefore they cannot be used for prediction and are disregarded.

As shown in Fig. S31b approximately 50% of cores are predictive, in stark comparison to BFS reference groups, with the random null model performing even worse.

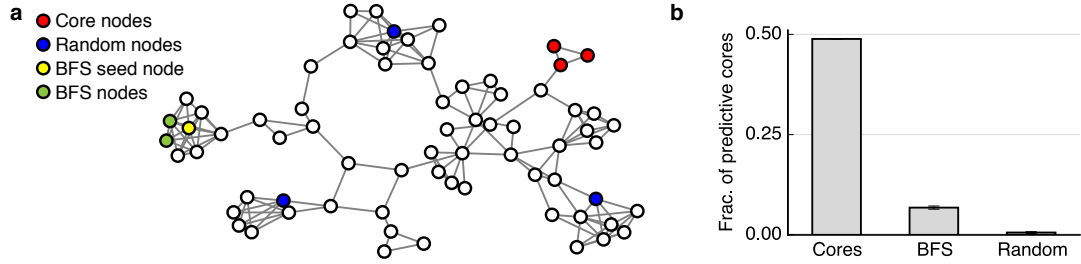


Figure S31: *Social prediction. a, Daily graph of interactions, illustrating cores and construction of null models. b Percentage of socially predictive cores within each category. A group is predictive if it at least once has correctly predicted the arrival of the missing individual. For the reference models errorbars are calculated across $n = 100$ independent trials.*

S7 Coordination of meetings

This section describes how we calculate the amount of coordination leading up to gatherings. Prior to a meeting individuals might need to coordinate about when and where to meet. This coordination can be conveyed through various means: (1) individuals can organize in real time through electronic means, such as online social networks and mobile phones, (2) they can verbally have scheduled meetings beforehand, i.e. at previous gatherings, (3) or attend routine driven pre-scheduled meetings, arranged by an institution, e.g. the university. We consider the coordination in the hours leading up to a meeting, measured in terms of increased calling and texting activity. Because calling frequencies change over the course of a day, and because individuals can have fundamentally distinct calling patterns [1, 37, 38], we compare activity leading up to a meeting to hour-by-hour dynamic individual baselines, defined as

$$c_t = \frac{1}{N} \sum_{n=1}^N \frac{a_t^n}{\widetilde{a_t^n}}, \quad (\text{S18})$$

where N is the number of individuals participating in the gathering, a_t^n is the activity of person n , and $\widetilde{a_t^n}$ is the baseline activity. The equation denotes increased coordination levels t hours before a meeting. Because gatherings have a broad distribution of lifetimes (Fig. S8), we restrict the calculation to individuals that participate in the first hour of the meeting. According to Fig. S32a-b, meetings during the weekend require more coordination than meeting during weekdays. This implies that weekend behavior is less scheduled, emphasizing the problem of predicting behavior using traditional routine-based measures.

In addition, Fig. S32c reveals that meetings, independently of size, require the same amount of coordination per person, illustrating the validity of Fig. S32a for varying gathering sizes.

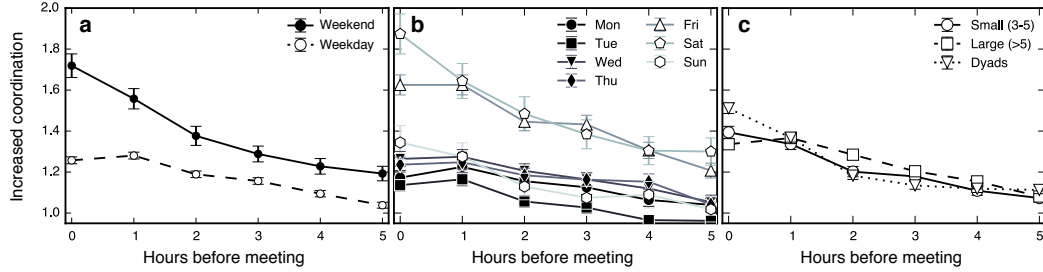


Figure S32: *Increased coordination prior to a meeting. Calculated for all nodes participating in the first hour of a gathering. a, Required amounts of coordination between nodes, depending on when the gatherings meets. More coordination is required to organize meetings during weekends (Friday 4 pm - Sunday) than during weekdays (Monday - Friday 4 pm). b, Further sub-dividing the categories from panel a reveals that Fridays and Saturdays are special. c, Effect of size of the meeting on coordination. On average it requires equal amounts of coordination, per person, to organize a meeting, independent of the size of the group.*

References

- [1] Stopczynski, A. *et al.* Measuring large-scale social networks with high resolution. *PLoS ONE* **9**, e95978 (2014).
- [2] Sekara, V. & Lehmann, S. The strength of friendship ties in proximity sensor data. *PLoS One* **9**, e100915 (2014).
- [3] Eagle, N., Pentland, A. S. & Lazer, D. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences* **106**, 15274–15278 (2009).
- [4] Clauset, A. & Eagle, N. Persistence and periodicity in a dynamic proximity network. *arXiv preprint arXiv:1211.7343* (2012).
- [5] Sulo, R., Berger-Wolf, T. & Grossman, R. Meaningful selection of temporal resolution for dynamic networks. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, 127–136 (ACM, 2010).
- [6] Krings, G., Karsai, M., Bernhardsson, S., Blondel, V. D. & Saramäki, J. Effects of time window size and placement on the structure of an aggregated communication network. *EPJ Data Science* **1**, 1–16 (2012).
- [7] Simmel, G. Quantitative aspects of the group. *The Sociology of Georg Simmel* 87–177 (1950).
- [8] Coser, L. A. & Merton, R. K. *Masters of sociological thought: Ideas in historical and social context* (Harcourt Brace Jovanovich New York, 1971).
- [9] Ward Jr, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* **58**, 236–244 (1963).
- [10] Gower, J. C. & Ross, G. Minimum spanning trees and single linkage cluster analysis. *Applied statistics* 54–64 (1969).
- [11] Newman, M. E. & Girvan, M. Finding and evaluating community structure in networks. *Physical review E* **69**, 026113 (2004).
- [12] Ahn, Y.-Y., Bagrow, J. P. & Lehmann, S. Link communities reveal multiscale complexity in networks. *Nature* **466**, 761–764 (2010).
- [13] Palla, G., Barabási, A.-L. & Vicsek, T. Quantifying social group evolution. *Nature* **446**, 664–667 (2007).
- [14] Reis, H. T. & Wheeler, L. Studying social interaction with the rochester interaction record. *Advances in experimental social psychology* **24**, 269–318 (1991).
- [15] Davis, A., Gardner, B. B. & Gardner, M. R. *Deep South; a social anthropological study of caste and class*. (University of Chicago Press, 1941).

- [16] Tibshirani, R., Walther, G. & Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**, 411–423 (2001).
- [17] Milligan, G. W. & Cooper, M. C. An examination of procedures for determining the number of clusters in a data set. *Psychometrika* **50**, 159–179 (1985).
- [18] Shannon, C. A mathematical theory of communication. *The Bell System Technical Journal* **27**, 379–423 (1948).
- [19] Song, C., Qu, Z., Blumm, N. & Barabási, A.-L. Limits of predictability in human mobility. *Science* **327**, 1018–1021 (2010).
- [20] Fano, R. M. *Transmission of information: a statistical theory of communications* (MIT Press, 1961).
- [21] Jensen, B. S., Larsen, J. E., Jensen, K., Larsen, J. & Hansen, L. K. Estimating human predictability from mobile sensor data. In *Machine Learning for Signal Processing (MLSP), 2010 IEEE International Workshop on*, 196–201 (IEEE, 2010).
- [22] Bagrow, J. P. & Lin, Y.-R. Mesoscopic structure and social aspects of human mobility. *PloS one* **7**, e37676 (2012).
- [23] Gonzalez, M. C., Hidalgo, C. A. & Barabasi, A.-L. Understanding individual human mobility patterns. *Nature* **453**, 779–782 (2008).
- [24] Trevisani, E. & Vitaletti, A. Cell-id location technique, limits and benefits: an experimental study. In *Mobile Computing Systems and Applications, 2004. WMCSA 2004. Sixth IEEE Workshop on*, 51–60 (IEEE, 2004).
- [25] Lin, M., Hsu, W.-J. & Lee, Z. Q. Predictability of individuals’ mobility with high-resolution positioning data. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, 381–390 (ACM, 2012).
- [26] Cuttone, A., Lehmann, S. & Larsen, J. E. Inferring human mobility from sparse low accuracy mobile sensing data. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, 995–1004 (ACM, 2014).
- [27] Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, vol. 96, 226–231 (1996).
- [28] Wrzus, C., Hänel, M., Wagner, J. & Neyer, F. J. Social network changes and life events across the life span: A meta-analysis. *Psychological bulletin* **139**, 53 (2013).
- [29] Lu, X., Bengtsson, L. & Holme, P. Predictability of population displacement after the 2010 haiti earthquake. *Proceedings of the National Academy of Sciences* **109**, 11576–11581 (2012).

- [30] Danon, L., Diaz-Guilera, A., Duch, J. & Arenas, A. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment* **2005**, P09008 (2005).
- [31] Lancichinetti, A. & Fortunato, S. Community detection algorithms: a comparative analysis. *Physical review E* **80**, 056117 (2009).
- [32] Eagle, N. & Pentland, A. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing* **10**, 255–268 (2006).
- [33] Crandall, D. J. *et al.* Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences* **107**, 22436–22441 (2010).
- [34] Cho, E., Myers, S. A. & Leskovec, J. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1082–1090 (ACM, 2011).
- [35] Wang, D., Pedreschi, D., Song, C., Giannotti, F. & Barabasi, A.-L. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1100–1108 (ACM, 2011).
- [36] De Domenico, M., Lima, A. & Musolesi, M. Interdependence and predictability of human mobility and social interactions. *Pervasive and Mobile Computing* **9**, 798–807 (2013).
- [37] Onnela, J.-P. *et al.* Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences* **104**, 7332–7336 (2007).
- [38] Jo, H.-H., Karsai, M., Kertész, J. & Kaski, K. Circadian pattern and burstiness in mobile phone communication. *New Journal of Physics* **14**, 013055 (2012).

Understanding scientific collaboration

SCIENTISTS often worry about their future. Where will our careers take us, which paths of research should we focus on, and how should we secure the required funding? As a consequence a lot of work has been devoted to quantifying science; from forecasting the future impact of publications (Wang et al., 2013), predicting the scientific success of individual researchers (Acuna et al., 2012; Sarigöl et al., 2014), to understanding the measures which we use to quantify impact (Lehmann et al., 2006). Contrary to popular belief of the individual genius, science in nowadays performed in teams (Wuchty et al., 2007). For example, de Solla Price (1963) examined in his famous book the change in team size in chemistry from 1910 to 1960 and projected that by the 1980's 0% of papers would be written by single authors. It did not go quite as foretold, nonetheless teams increasingly dominate single author publications. In order to gain a deeper understanding of science we, therefore, need to view research from the perspective of groups. As such, recent studies have investigated how collaborations arise and grow (Milojević, 2014), how teams succeed (Klug and Bagrow, 2014), and how credit should be allocated between team members (Shen and Barabási, 2014).

Within multi-author publications there exists clear signs of mentorship (Malmgren et al., 2010; Callaway, 2015). However, we still lack a quantitative understanding of such chaperone bonds and, in general, how knowledge is passed down by between different generations of scientists. Paper IV presents a study where we perceive teams as conduits through which knowledge can be transferred between researchers. To quantify the chaperone phenomenon we examine how scientists transition from junior to senior status within the same scientific journal. Specifically, we ask whether the principal investigator (PI) of a paper has previously published within the same venue as a junior author; if so, is the PI of a paper more inclined to be listed again in the same role in future publications within the same journal? Our results show that the magnitude of the chaperone phenomenon varies depending on the branch of science. It is pronounced for medical and biological sciences, more subtle for chemistry and physics, and in the case of mathematics there is no clear effect of mentorship. Furthermore, having previously been listed as a junior researcher in one of the multidisciplinary journals such as *Nature*, *Science*, and *PNAS*, significantly enhances your chances of being listed as PI. These findings document that scientific training plays a fundamental role in acquiring the necessary experience, expertise and skills to publish in specific venues.

The chaperone phenomenon in science

Vedran Sekara¹, Roberta Sinatra^{2,3}, Pierre Deville^{2,4}, Sebastian Ahnert⁵, Albert-László Barabási^{2,3,6} & Sune Lehmann^{1,7}

¹*Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark,*

²*Center for Complex Network Research, Northeastern University, Boston, USA,*

³*Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, USA,*

⁴*Department of Applied Mathematics, Université catholique de Louvain, Belgium,*

⁵*Theory of Condensed Matter, Cavendish Laboratory, University of Cambridge, United Kingdom,*

⁶*Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, USA.*

⁷*The Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark.*

In science, anecdotal evidence exists that high-achievers are often protégés of illustrious mentors ¹⁻³. While projects like the mathematical genealogy document clear signs of such chaperone bonds between renowned scientists ⁴, we lack systematic quantitative evidence of the role of apprenticeship in scientific publishing and, in general, of how scientific knowledge is passed down between different generations of scientists ^{4,5}. Here we quantify the chaperone phenomenon by considering how scientists transition to senior status within the same scientific journal. We show that a scientist is unlikely to appear as senior author if she has not published already in the same journal, and that this trend has become more pronounced in the last decade. We illustrate that the chaperone phenomenon has different magnitude for journals that belong to different branches of science, being more pronounced for medical and biological sciences but more subtle for the natural sciences. These findings document the fundamental role played by scientific training to acquire the necessary experience, expertise and skills to publish in specific venues.

In a letter addressed to Robert Hooke (1676), Isaac Newton famously stated: “If I have seen further it is by standing on the shoulders of giants.”, indicating that his achievements have been possible only thanks to solid scientific foundations built and passed down to him by his predecessors. Evidence of the fundamental role played by mentorship in science ⁵ is offered by the dynamics of scientific multi author publications. Indeed, high impact works are often performed by collaborative teams ^{6,7}, whose composition is usually heterogenous in terms of experience and age, with case studies showing that experience ^{8,9} and leadership ¹⁰ are important factors for achieving success. Moreover, a recent study has highlighted that junior researcher tend to work on more innovative topics but that they are in need of mentorship ^{11,12}. Thus teams are perfect conduits for senior scientists to train younger colleagues. To quantify the chaperone effect in science, we examine how the process manifests itself in multi-authored publications. We consider the publication history of researchers within a specific scientific journal (see SI for the author name disambiguation procedure), asking: has the principal investigator (PI) of a paper published previously in the same venue as a junior author? If so, is the PI of a paper more inclined to be listed again in the same role in future papers in the same journal?

We consider 6.4 million papers published between 1960 and 2012 in 394 scientific journals; covering the scientific fields of mathematics, physics, chemistry, biology, and medicine. In addition, we also include the top 3 multidisciplinary journals: Nature, Science, and PNAS. Only papers that are labeled as academic in the ISI Web of Knowledge (Thomson Reuters) database are taken into account. In addition, to make sure we are only dealing with original research we disregard papers that have titles containing the terms: *comment*, *reply*, or *retracted article*. In our analysis,

we assume that the PI is always listed last in a paper's author list, a common practice in many scientific fields^{13,14}. Note, however, that our analysis is not affected if the author list of some papers does not mirror seniority roles, as in the case of the alphabetical order. For each journal we divide PIs into 3 categories: *new* denotes authors who have never published previously in that specific journal, *non-last* are authors who have appeared before but only as junior (non-last) authors, and finally *last* denotes authors previously listed as last author in the journal.

Figure 1 shows the fraction over time of the three categories of last-authors for the scientific journal Nature. From 1990 to year 2000 the fraction of PIs that represent each category is largely constant and last authors with no experience of publishing in the journal and last authors with previous last-author status are equally represented, and their number is twice that of junior authors. Starting in year 2000, however, the fraction of PIs that are new to the journal, $p(\text{new})$, rapidly declines to the advantage of last authors with previous experience as last authors. Thus, it becomes increasingly rare to publish as the senior author in Nature without previous publishing experience in the journal. To quantify the chaperone phenomenon, we study the relative magnitude of the *non-last*→*last* in respect to *new*→*last* transition. To capture a genuine effect, we need to compare the observed transitions with those occurring in a system where the ordering of author names is not relevant¹⁵. Therefore, we study the ratio $c = p(\text{non-last}) / p(\text{new})$ and compare it to c_{random} , the ratio obtained in a null model where we randomly permute the order of author names in each paper. We call c/c_{random} the of magnitude the chaperone effect, and the chaperon phenomenon occurs when $c/c_{\text{random}} > 1$, i.e. when the transition *non-last*→*last* is more frequent than *new*→*last* in a statistical significant way, capturing the tendency of authors to become PIs after having published

as junior authors before. Having witnessed and gone through the entire publication process once, increases the chances of publishing in similar journals again, since the author is familiar with their inner workings.

Next we ask whether the amount of experience acquired through chaperone bonds differs between scientific fields. First, we find that the chaperone phenomenon, within each discipline, is stable over time (Figure 2), indicating that a constant amount of experience and training is required to transition between junior and senior status. Because the level of apprenticeship is fairly stable, we can collapse the distributions to quantify differences between fields. Figure 3 shows that the chaperone phenomenon is pronounced for medical and biological sciences, while it is more subtle for physics and chemistry. In the case of mathematics there is no clear sign that mentorship influences the transition between junior and senior levels. In fact there is a 114% difference between the chaperone phenomena for mathematics and biology, reflecting the way science is performed within each discipline. Incorporating multidisciplinary journals we see a clear relationship between having published in them as a junior researcher and the probability of publishing in them as PI, clearly illustrating that experience provided by appropriate mentors is a considerable factor of transitioning between junior and senior levels in these high impact journals. While we cannot pinpoint which facet of experience that is most important to succeed, surrounding a young scientist with experienced researchers will overall have a positive impact on her career.

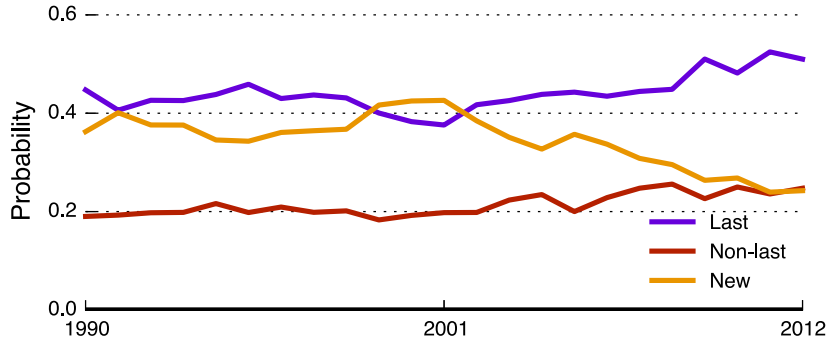


Figure 1: **Probability of being listed as PI in Nature given previous publication history.** The last authors of all papers published each year in Nature are divided into three categories: *new* authors, that have never published in Nature before, *non-last* authors, that have published in Nature before only at junior level, and *last* authors, that have already previously published as last authors.

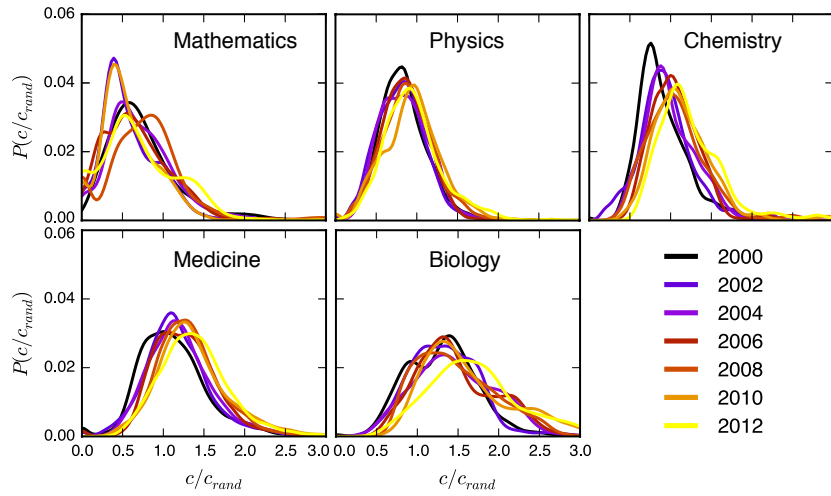


Figure 2: **Chaperone distributions for various fields as function of time.** Dividing journals into branches of science reveals that each field has a specific signature.

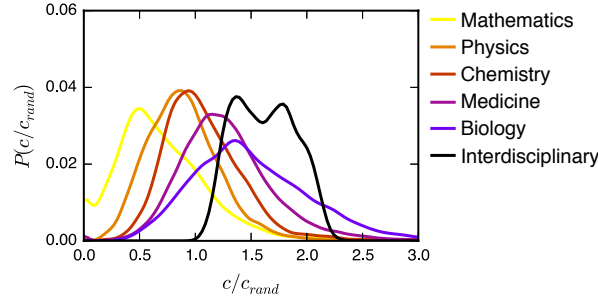


Figure 3: **Comparison of chaperone phenomenon between scientific fields.** Distributions for the past 12 years are collapsed into single distributions and enable us to compare separate scientific fields. For the different disciplines we find $\langle c/c_{rand} \rangle_{\text{math}} \simeq 0.71$, $\langle c/c_{rand} \rangle_{\text{physics}} \simeq 0.90$, $\langle c/c_{rand} \rangle_{\text{chemistry}} \simeq 1.08$, $\langle c/c_{rand} \rangle_{\text{medicine}} \simeq 1.28$, and $\langle c/c_{rand} \rangle_{\text{biology}} \simeq 1.52$, while the effect for interdisciplinary journals is $\langle c/c_{rand} \rangle_{\text{interdisc}} \simeq 1.61$. A Wilcoxon rank sum test, moreover, illustrates that the distributions are distinguishable $p \ll 0.05$.

References

1. <http://genealogy.math.ndsu.nodak.edu/>.
2. http://en.wikipedia.org/wiki/Academic_genealogy_of_theoretical_physicists.
3. Parberry, I. & Johnson, D. S. *The SIGACT theoretical computer science genealogy: Preliminary report* (2004).
4. Malmgren, R. D., Ottino, J. M. & Amaral, L. A. N. The role of mentorship in protégé performance. *Nature* **465**, 622–626 (2010).
5. Chao, G. T., Walz, P. & Gardner, P. D. Formal and informal mentorships: A comparison on mentoring functions and contrast with nonmentored counterparts. *Personnel psychology* **45**, 619–636 (1992).
6. Wuchty, S., Jones, B. F. & Uzzi, B. The increasing dominance of teams in production of knowledge. *Science* **316**, 1036–1039 (2007).

7. Milojević, S. Principles of scientific research team formation and evolution. *Proceedings of the National Academy of Sciences* **111**, 3984–3989 (2014).
8. Delmar, F. & Shane, S. Does experience matter? the effect of founding team experience on the survival and sales of newly founded ventures. *Strategic Organization* **4**, 215–247 (2006).
9. Klug, M. & Bagrow, J. P. Understanding the group dynamics and success of teams. *arXiv preprint arXiv:1407.2893* (2014).
10. Johnstone, R. A. & Manica, A. Evolution of personality differences in leadership. *Proceedings of the National Academy of Sciences* **108**, 8373–8378 (2011).
11. Packalen, M. & Bhattacharya, J. Age and trying out new ideas, working paper. *National Bureau of Economic Research* **20920** (2015).
12. Callaway, E. Young scientists go for fresh ideas. *Nature* **518**, 283–284 (2015).
13. van Dijk, D., Manor, O. & Carey, L. B. Publication metrics and success on the academic job market. *Current Biology* **24**, R516–R517 (2014).
14. Austin, J. What it takes. *Science* **344**, 1422 (2014).
15. Sinatra, R., Wang, D., Deville, P., Song, C. & Barabási, A.-L. Quantifying patterns of scientific excellence. *in review* (2015).

Acknowledgements Put acknowledgements here.

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence and requests for materials should be addressed to S.L. (email: sunelehmenn@gmail.com).