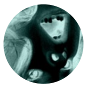
Measuring Large-Scale Social Networks with High Resolution

Article *in* PLoS ONE · April 2014

DOI: 10.1371/journal.pone.0095978 · Source: PubMed

7 authors, including:



my Madsen
University of Copenhagen

Jakob Eg Larsen
Technical University of Denmark

Measuring large-scale social networks with high resolution

Arkadiusz Stopczynski[1,*], Vedran Sekara[1], Piotr Sapiezynski[1], Andrea Cuttone[1], Jakob Eg Larsen[1], Sune Lehmann[1,2]
1 Technical University of Denmark, 2 The Niels Bohr Institute
∗ **E-mail: arks@dtu.dk**

WORKING PAPER

Contents

1 Abstract

This paper describes the deployment of a large study designed to collect data about human interactions with unprecedented temporal resolution and depth, both in terms of duration and number of distinct channels. To collect this data, we use state-of-the-art smartphones as social sensors. We collect data on face-to-face interactions, telecommunication, social networks, geolocation, and background (personality, demographic, health, politics) information for a densely connected and coherent population of 1,000 individuals. From a basic science perspective, our motivation is to advance the understanding of highly connected social systems. Building from our densely sampled system, we are interested understanding what working on subsamples implies for results on social network analysis; we explore this topic by comparing single-channel information with our layered picture of interactions. In terms of network science, we aim to expand our understanding of the dynamics of the social network, how it changes over time, and the causes and implications of the processes in play. Further, we want to understand how information—in various forms—spreads in such dynamic networks, to better model contagion, influence, and cooperation. Underlying the basic science agenda, is the technical component: custom energy-efficient data collection software running on the phones, database design needed to store the collected data, research and development focused on analyzing and interpreting the signals emerging from the phones in terms of human social interactions. The final area of interest addressed is privacy in such massive sensor-driven data collection from human participants, understood as tools, methods, and philosophy designed to make the data useful and sharable without compromising the privacy of the study participants. We go over motivation, results, challenges, and details on the initial deployments this study (SensibleDTU).

2 Introduction

Driven by 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. In part this is 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 is '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 (Motivation 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 which 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 user, for example based on call detail records (CDR) studies [4] or Twitter analysis [5]. Here, we are interested in capturing deeper data, that is, looking at multiple channels from sizable populations. Using big data collection and analysis techniques, which can scale in number of users, 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 for a large, highly connected population. We also aim to create a data set collected in an interactive way, where we can change the collection process. This enables us to rapidly adapt and change our collection method if current data, for example, have the wrong resolution for a certain question we would like to answer. We have designed our data collection setup in such a way we are able to deploy experiments, since we know that causal inference is notoriously complicated in network settings [6], and perform continuous quality control of the data collected. The mindset of real-time data access can be extended beyond pure research, were the goal is to monitor data quality and perform 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 to approach the data in the same way large internet services do—as a resource that can be manipulated and made available in real-time, since it inevitably loses value over time.

In order to realize the interactive data collection described above, 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 changes 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, working 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 mankind is moving in this direction: 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 living with highly detailed lifelog, such as every single electronic message they have ever exchanged with their friends. The status quo is we need to actively opt out for our experiences not to be 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 explicitly turn on private mode.

Human lives, especially when seen in the perspective of months and years, take place in multiple dimensions. Capturing only a single channel, even for the entire persons' life, limits the expertise 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 scientists are all interested in distinct questions, and traditionally use very different methods. We believe that it is when those groups start working together, qualitatively better findings can be made. Today we have the tools to collect data that can allow us to begin bridging some of the divisions across scientific fields.

In this paper we describe a large scale study observing student lives through multiple channels: SensibleDTU. With an iterative approach to deployments, it is an example of an interdisciplinary approach. We collect data from multiple sources, including questionnaires, online social networks, and smartphones handed out to students. Data from all those channels create multi-layered view of the individuals, their networks, and their environments; those views can be then examined separately and jointly by researchers of different expertise. We are building SensibleDTU as a framework for long-lived extensible studies, and the 2013 deployment of SensibleDTU has been designed as part of a close collaboration with researchers from the social sciences, natural sciences, medicine (public health), and humanities (see Acknowledgements for details). Currently in the second iteration, we have deployed phones for about 1,000 participants, growing a dataset of unprecedented size and resolution. In addition to the core task of collecting the deep behavioral data, we also experiment with creating rich services for our participants and improving privacy.

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 the implemented privacy and informed consent practices, 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 the 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 researches of all fields to contact the authors for the purpose of defining novel projects around the SensibleDTU testbed.

3 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—phone call records, GPS traces, credit card transactions, webpage visits, emails, data from social networks [9]. CSS focuses on questions that can now be studied using data-driven computational analyses of data sets such as the ones mentioned above, and which could only previously be addressed self-reported data or direct observations: dynamics in work groups, face-to-face interactions, human mobility, or information spread. The hope is that such a data-driven approach will bring new types of insight, not available from classical 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 the central methods of data collection, provide a brief overview of results obtained from the analysis of CSS data, and mention some principles regarding privacy and data treatment.

3.1 Data collection

Many of the CSS studies have been performed on call detail records (CDRs)—records of user 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 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], sociometetric badges [13], as well as smartphones for the data collection [14–17]. 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 [18]. Collecting data with smartphones presents, however, several limitations: sensing is mainly limited to pre-installed sensors, which may not be of highest quality, and 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 [19], SocioXensor [20], MyExperience [21], Anonysense [22], CenceMe [23], Cityware [24], Darwin phones [25], Vita [26], and ContextToolbox [27].

Running longitudinal rich behavioral data collection from large populations presents multiple logistical challenges and only few studies have attempted it so far. In the Reality Mining study, datasets from 100 mobile phones was collected for nine months [28]. In the Social fMRI study, 130 participants carried smartphones running the Funf mobile application [29] for 15 months [30]. Data was also collected from Facebook, credit card transactions, and surveys were pushed to the participants' phones. The Lausanne Data Collection Campaign [31] [32] featured 170 volunteers in the Lausanne area of Switzerland, between October 2009 and March 2011. In the SensibleOrganization study [33] 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 with 20 participants from a large university campus have been so far reported from the OtaSizzle study [34], and in the Locaccino study [35], location within a metropolitan region was recorded for 489 users for varying periods, ranging from seven days to several months.

3.2 Data analysis

Below we provide selected examples of results obtained from analysis of CSS datasets in various domains.

3.2.1 Human Mobility

Gonzales et al. analyzed six months of CDRs of 100 000 users, finding that human mobility is quite predictable, with high spatial and temporal regularity, and few highly frequented locations [36]. This result was extended by Song et al., where three months of CDRs from 50 000 individuals were analyzed, finding a 93% upper bound of predictability of human mobility, a number which applies to most users regardless of different travel patterns and demographics [37]. 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 [38]. Bagrow et al. analyzed 34 weeks of CDRs for 90,000 users, identifying habitats—groups of related places—and finding that most individuals in their dataset have between 5 and 20 habitats [39]. De Domenico et al. showed in [40] 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.

3.2.2 Social Interactions

Face-to-face interactions can be used to model social ties over time and organizational rhythms in response to events [28, 41, 42]. Comparing those interactions with Facebook networks, Cranshaw et al. found that meetings in locations of high entropy (featuring diverse set of visitors) are less indicative than meetings in locations visited by a small set of users [35]. Clauset et al. found that a natural time scale of face-to-face social networks is 4 hours [43].

Onnela et al. analyzed CDRs from 3.9 millions users [44], finding evidence supporting the weak ties hypothesis [45]. Lambiotte et al. analyzed CDRs from 2 millions user, finding that the probability of the existence of the links decreases as d^{-2}, where d is the distance between users [46]. In another study with CDRs from 3.4 million users, the probability was found to decrease as $d^{-1.5}$ [47]. Hidalgo et al. found—analyzing CDRs for 2 millions users—that persistent links tend to be reciprocal and associated with low degree nodes [48].

Miritello et al. analyzed CDRs for 20 millions people, and observed that individuals have a finite limit of number of active ties, and two different strategies for social communication [49, 50]. Sun et al. analyzed 20 million bus trips from 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 [51].

3.2.3 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 epicenter, and determination of regions requiring relief efforts [52]. Aharony et al. performed and evaluated a fitness activity intervention with different reward schemes, based on face-to-face interactions [30] while Madan et al. studied how different illnesses (common cold, depression, anxiety) manifest in common mobile-sensed features (WiFi, location, Bluetooth) and the effect of social exposure on obesity [53].

3.2.4 Influence and Information Spread

Chronis et al. [15] and Madan et al. [54] 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 subnetworks [55]. Aharony et al. analyzed the usage of mobile apps in relation with face-to-face interactions, finding that more face-to-face interaction increases the number of common applications [30]. Using RFID for sensing face-to-face interactions, Isella et al. estimated the most probable vehicles for infection propagation [56]. With a a similar technique, but applied to 232 children and 10 teachers in a primary school, Stehle et al. described a strong age homophily in children interactions [57].

Bagrow et al. showed how CDRs 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 [58]. Karsai et al. analyzed CDR from six millions users to find that strong ties tend to constrain the information spread within localized groups of individuals [59].

3.2.5 Socioeconomics and Organizational Behavior

Face-to-face contacts and email communication for employees in a real work environment can be used to predict job satisfaction and group work quality [33]. Having more diverse social connections is correlated with economics opportunities, as found in the study containing CDRs of over 65 millions users [60]. 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 [30]. Language regions in a country can be identified based solely on CDRs, as showed in a study of Belgian users [61].

3.3 Privacy

Data collected about human participants is sensitive and ensuring privacy of the users is a fundamental requirement—even when those users may have limited understanding of the implications of data sharing [62, 63]. A significant literature exists regarding the understanding the possible attacks that can be performed on personal data, such as unauthorized analysis of personal data [64] to decode daily routines [65] or friendships [41] of the participants. In *side channel information* attacks data from public datasets (e.g. online social networks) are used to re-identify users [66–68]. Even connecting different records of one user within the same system can compromise privacy [66]. Specific attacks are also possible in network data, as nodes can be identified based on the network structure and attributes of the neighbors [69, 70].

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, PIIs can be completely removed, hidden by aggregation, or transformed to be less identifiable, resulting in a trade-off between privacy and utility [71]. Substituting PII with the correspondent one-way hash allows to remove plaintext information and break the link to other datasets. This method, however, does not guarantee protection from re-identification [72–74]. $K-$anonymity [75] is a technique of ensuring that it is not possible to distinguish any user from at least $k-1$ other in the dataset; studies have shown that this method may be often too weak [65]. $L-$diversity [76] and $t-$closeness [77] 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 [78–82]. *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 [83–86]; this technique has been applied, for example, to vehicle positioning data [87] and medical records [88].

The flows of data—creation, copying, sharing—can be restricted. *Information Flow Control* solutions such as [89–91] attempt to regulate the flow of information in digital systems. *Auditing* implementations such as [92–94] 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 [95–98]. *Watermarking* identifies records using hidden fingerprints, to allow traceability and identify leaks [99–101].

4 Motivation

Here we describe the primary motivations for deploying the SensibleDTU study, featuring deep and high-resolution data and a longitudinal approach.

4.1 Multiplexity

A majority of big data studies have been done with datasets containing data from a single source, such as call detail records (CDRs) [4], RFID sensors [102], Bluetooth scanners [103], or online social networks activity [2]. Although, as we presented in the Related Work section, analyzing these datasets has led to some exciting findings, however, we may 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, calling each other, meeting face-to-face, or exchanging emails. Observing only one channel may introduce bias that is difficult to estimate [11]. Ranjan et al. investigated in [104] how CDR datasets—containing samples dependent upon user activity and requiring user participation—may bias our understanding of human mobility. The authors used data activities as 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 [105], they found that the CDRs reveal only a small fraction of users' mobility, when compared to data activity. The identified home and work locations—considered the most important ones—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, datasets are obtained from different populations and in different moments, it is difficult to cross-validate or combine findings, and 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, utilizing the varied expertises and results they generate to provide 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 [106, 107] begin to explore the topic, a coherent theory describing multiplex, weighted, and directed networks remains beyond the frontier of our current understanding.

4.2 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 once a week or less [108]. Further, CDRs are typically provided by a single provider with some finite market share. If the market share is 20% of the population and you consider only links internal to your dataset, that translates to only 4% of the total number of links, assuming random network and random sampling [4]. Thus, while CDRs might be sufficient for analyses 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 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 automatically measuring Bluetooth proximity, recording location, and visible WiFi networks [9, 28, 30]. In cases where we cannot observe users passively or when something simply goes wrong with the data collection, we aim to use the redundancy in the channels: if the user turns off Bluetooth for some period, we can still estimate the proximity of users 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 users to include in the analysis. As investigated in [104], when only high-frequency voice-callers are chosen from a CDR dataset for the purpose of analysis, it can incur biases in Shannon entropy values of mobility. Similarly, as shown in [108], 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 users, we risk introducing the 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 users, the SensibleDTU population is also a subset, and every network we analyze will be sampled in some way, see [109] for a review on sampling. While SensibleDTU 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* [110].

4.3 Dynamics

The networks and behaviors we observe are not static, displaying 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, 111]. The time scale of the changes varies and depends on many factors: semester cycle in students life, changing schools or work, or simply getting older, to name just a few. Without following such dynamics, focusing on a single temporal slice, we risk missing an important aspect of human nature. To capture it, we need long-term studies, following the participants through the period of months and years.

But our behavior is not static in even very short intervals either. We have daily routines, meeting with different people in the morning and hanging out with other in the evening, see Figure 4. Our workdays may see us going to places and interacting with people differently than on weekends. It is easy to miss such dynamics when the quality of the data is insufficient, either not sampled frequently enough or 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 [112], but 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.

4.4 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 the 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 observed 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 do distinct stimuli, addressing the topic of causality as well as confounding factors both of which have proven problematic [113, 114] for the current state of the art [115, 116].

Simultaneously, we monitor the data quality not only on the most basic level of a user—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 students we interviewed reported turning the 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 those users, 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 [117]. Applications for participants create value, increase engagement, and may even allow to deploy studies without buying a large number of smartphones and provide them to participants. Our initial approach has included the development and deployment of a mobile app, which provides feedback about personal mobility and social interactions based on personal SensibleDTU data [118]. Preliminary results from the deployment, participants surveys, and usage logs suggest an interest in such applications, with a subset of participants repeatedly using the mobile app for personal feedback [119]. It is clear that feedback can potentially influence the study results: awareness of a certain behavior may cause participants to wish 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.

4.5 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 spread and the interplay between the various communication channels [120, 121].

A crucial step when studying the structure and dynamics of networks is to identify communities (densely connected groups of nodes) [122, 123]. In social networks, communities roughly correspond to social spheres. Recently we pointed out [124] that communities in many real world network display *pervasive overlap*, where each and every node belongs to more than one group. 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 which detect disjoint communities have operated successfully since the notion of graph partitioning was introduced in the 1970's [125], 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 disjoint [126]. 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.

5 Data Collection

The SensibleDTU study addresses the problem of single-modality data by collecting information from a number 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 [127]; finally, the smartphones carried by all participants record their location, tele communication 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 in 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 the SensibleDTU 2012 and 2013 deployments.

5.1 Data Sources

The data collected in the two studies was obtained from questionnaires, Facebook, mobile sensing, anthropological observations, and the wifi system on campus.

5.1.1 Questionnaires

In 2012 we deployed a survey containing 95 questions, covering socioeconomic factors, participants' working habits and Big Five Inventory (BFI) [128]. The questions were presented as a Google Form and participation in the survey was optional.

In 2013 we posed 310 questions per participant, prepared by a group of selected public health researchers, psychologists, anthropologists, and economists (see Acknowledgements).

The questions in the 2013 deployment included BFI, Rosenberg Self Esteem Scale [129], Narcissism NAR-Q [130], Satisfaction With Life Scale [131], Rotters Locus of Control Scale [132], UCLA Loneliness scale [133], Self-efficacy [134], Cohens perceived stress scale [135], Major Depression Inventory [136], and Panas [137], as well as health-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, branching depending on user's answers, and saving each individual's progress. The application is available as an open source project at `https://github.com/MIT-Model-Open-Data-and-Identity-System/SensibleData-Apps-Questionaires`. Participation in the survey is required for taking part in the experiment. In order to track and analyze temporal development, the survey (in a slightly modified form) is repeated every semester on all participating students.

5.1.2 Facebook

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

5.1.3 Mobile Sensing

For the data collection from mobile phones we used a modified version of the Funf framework [30] in both deployments. The data collection app 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 `https://github.com/organizations/OpenSensing`.

In the 2012 deployment, we manually kept track of which phone was used by eacg 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 are registered in the system by the students in an OAuth2 authorization flow initiated from the phone; the data are identified by a token stored on the phone and embedded in the data files. The sensed data are saved as locally encrypted sqlite3 databases and then uploaded to the server every 2 hours, provided the phone is connected to WiFi. Each file contains 1 hour of user data from all probes, saved as a single table. When uploaded, the data is decrypted, extracted, and included in the main study database.

5.1.4 Anthropological Observation

An anthropological observation study is included in the 2013 deployment. All participants have been informed about the role of the anthropologist, who participates in all student activities and courses, including group work, parties, and trips. The goal is to experience how connections are formed between people who did not know each other prior to the study. This also included the emergence of social 'logic' and rules in the newly created groups.

As part of the process of participant observation fieldwork, the anthropologist collects qualitative data regarding the formation of the social networks and the participant activities. We aim to relate these observations about the social ties forming and dissolving to the quantitative data obtained from the mobile sensing, creating a new approach for observing network dynamics and other social social phenomena. We are particularly interested in the relationship between the observations made by the embedded anthropologist and the data recorded using questionnaires and mobile sensing, answering questions such as 'what are the elements that are difficult to capture using our high resolution approach?'

and similarly from the perspective of the anthropologist 'what are the events that may be captured using technology that are invisible even to the trained observer?'. On a practical level, the underlying social rules governing the network formation are not necessarily possible to infer from a purely quantitative perspective; we can describe and model them, but still risk not fully grasping the real-world interactions that the model attempts to describe. The presence of an actual observer helps ground the mathematical modeling process.

5.1.5 Campus WiFi

For the 2012 deployment, between August 15, 2012, and May 22, 2013, we had been granted access to the campus WiFi system. Every 10 minutes the system provided data 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 users with students in our study. Campus WiFi data was not collected for the 2013 deployment.

5.2 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 is designed and implemented as an extensible framework for data collection, sharing, and analysis.

5.2.1 SensibleDTU 2012

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 mobilse 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.

5.2.2 SensibleDTU 2013

The 2013 system is built as an open Personal Data System (openPDS) [138] in an extensible fashion. The architecture of the system is depicted in Figure 1 and consisted of three layers: platform, services, applications. In the platform layer, the components common for multiple services were grouped, involving identity provider and user-facing portal for granting authorizations. The identity provider is based on OpenID 2.0 standard and enabled single sign-on (SSO) for multiple applications. The authorizations are realized using OAuth2 and can be used with both web and mobile applications. Users enroll into each study by accepting informed consent and subsequently authorized application to submit and access data from the study. The data storage is implemented using MongoDB. Users can see the status and change their authorizations on the portal site, the system included an implementation of the Living Informed Consent [3].

5.3 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.

5.3.1 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 to a few selected majors that had sufficient number of students interested in participating in the experiment. Directly asking students about their interest in SensibleDTU 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 the human nature by staging a competition, running for two weeks from the start of the semester. All students had access to a web forum, separate for each major, where they could post ideas that could be realized by the data we would collect, and subsequently vote on their own ideas or three seed ideas, which we provided. The goal of the competition was twofold, first we wanted students to register with their Facebook account 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 of the project and competition by the Dean in person and at one of 15 talks given—one to 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 behind at least one idea and were chosen to participate in the experiment.

The physical handout 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 make participant take better care of the phones, compared to if they had received them for free [139]. 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.

5.3.2 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 links on the pamphlet were promised phones before the start of their studies. Students from both branches of undergraduate programs were invited to participate (approximately 1500 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—accordingly, the informed consent and the questionnaire were a part of the registration. Again we required a symbolic cash deposit for each phone [139]. We pre-installed custom software on each phone to streamline the handout process; students still had to properly set up the phones (making them Bluetooth-discoverable, turning on WiFi connection, etc.).

For researchers considering similar projects with large scale handouts we provide the following recommendations: Engage the pool of subjects 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 amount can add up to significant sums, and must be handled properly.

6 Privacy

When collecting data of very high resolution, over an extended period, from a large population, it is crucial to properly address the privacy of the participants. We measure the privacy as a difference between what the user understands and consents to regarding her data and what in fact happens to those 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, making the user understand the situation, and to control, aligning the situation with the user's preferences. There is a tight loop where these tools interact: as the user 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 user expresses Living Informed Consent as described in [3].

Not all students are interested in privacy, in fact our observations were quite the opposite. In the SensibleDTU 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 users 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 the privacy-related issues.

In the 2012 deployment we used a basic informed consent procedure with an online form accepted by the users, 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 user. 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 [140] with English translation available at [141].

In the 2013 deployment, we use our backend solution (described in Section Backend System) to address the informed consent procedure and privacy in general. The account system, realized as an OpenID 2.0 server, allows 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 user data, all the applications identify users based only on the pseudonym identifiers. The applications can also access a controlled set of identity attributes for the purpose of personalization (e.g. greeting the user by name), subject to user OAuth2 authorization. The enrollment into the study, after the user 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 1). The acceptance of the document was recorded in the database by storing the username, timestamp, hash of the text presented to the user, as well as the git commit identifying the version of the form.

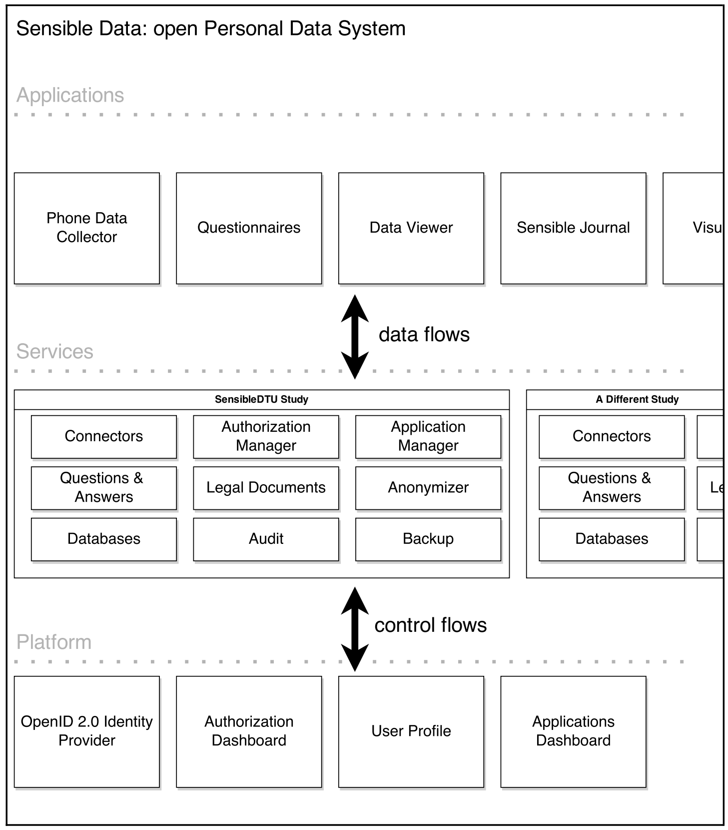


Figure 1. Sensible Data openPDS architecture. This system is used in the 2013 deployment and consists of three layers: platform, services, and applications. The platform contains element common for multiple services (in this context: studies). The studies are the deployments of particular data collection efforts. The applications are OAuth2 clients to studies and can submit and access data, based on user authorizations.

All the communication in the system is realized over HTTPS and endpoints were protected with short-lived OAuth2 bearer tokens. The text of the documents, including informed consent is stored in a git repository, allowing us to modify anything, while, importantly, still maintaining the history and being able to reference which version each user has seen and accepted. A single page overview of the status of the authorizations, presented in Figure 2, is an important step in moving beyond lengthy, incomprehensible legal documents accepted by the users blindly and giving more control over permissions to the user.

In the 2013 deployment, the users can 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 3, 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 *question & answer* framework, where the high-level features are extracted from the data before leaving the server, promoting better user understanding of data flows.

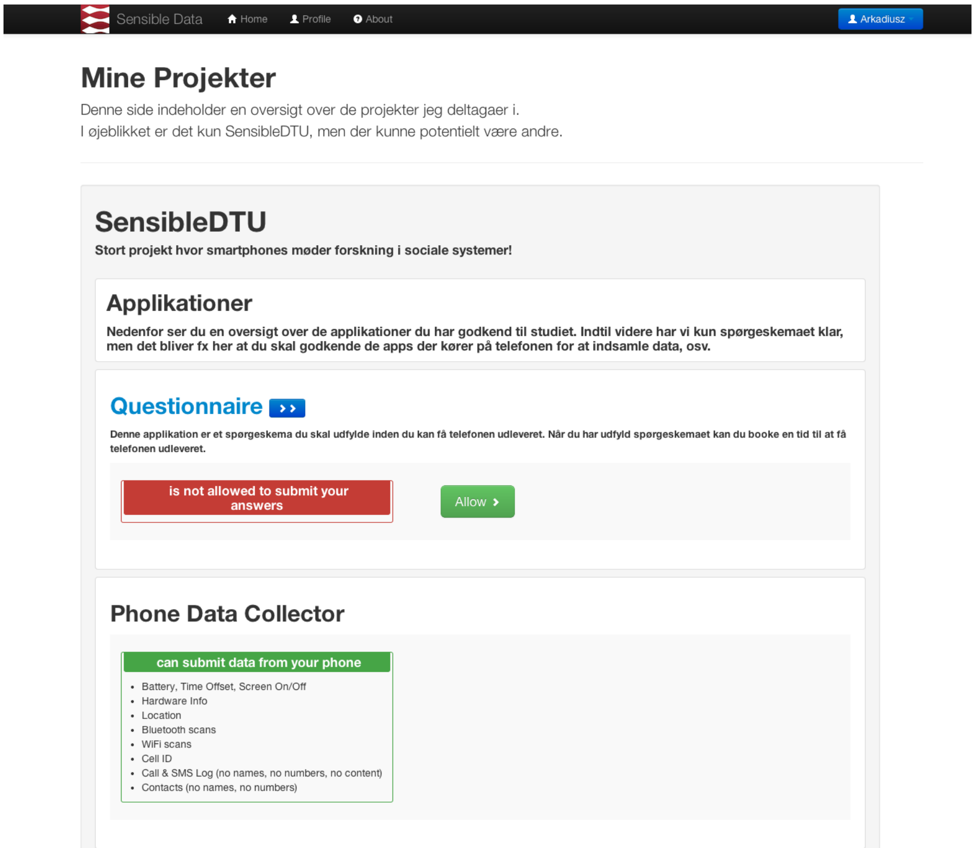


Figure 2. Authorizations page. Participants have an overview of the studies they are enrolled into and which applications that are able to submit to and access their data. This is an important step towards users' understanding what is happens with their data and to exercise control over it.

This is aligned with the vision of the open Personal Data Store [138].

Finally, for the purposes of engaging the users 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`).

7 Results

As described in the previous sections, the SensibleDTU 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 longitudinal nature of the data available to date.

7.1 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, handsfree headsets, tablets, and other wearables.

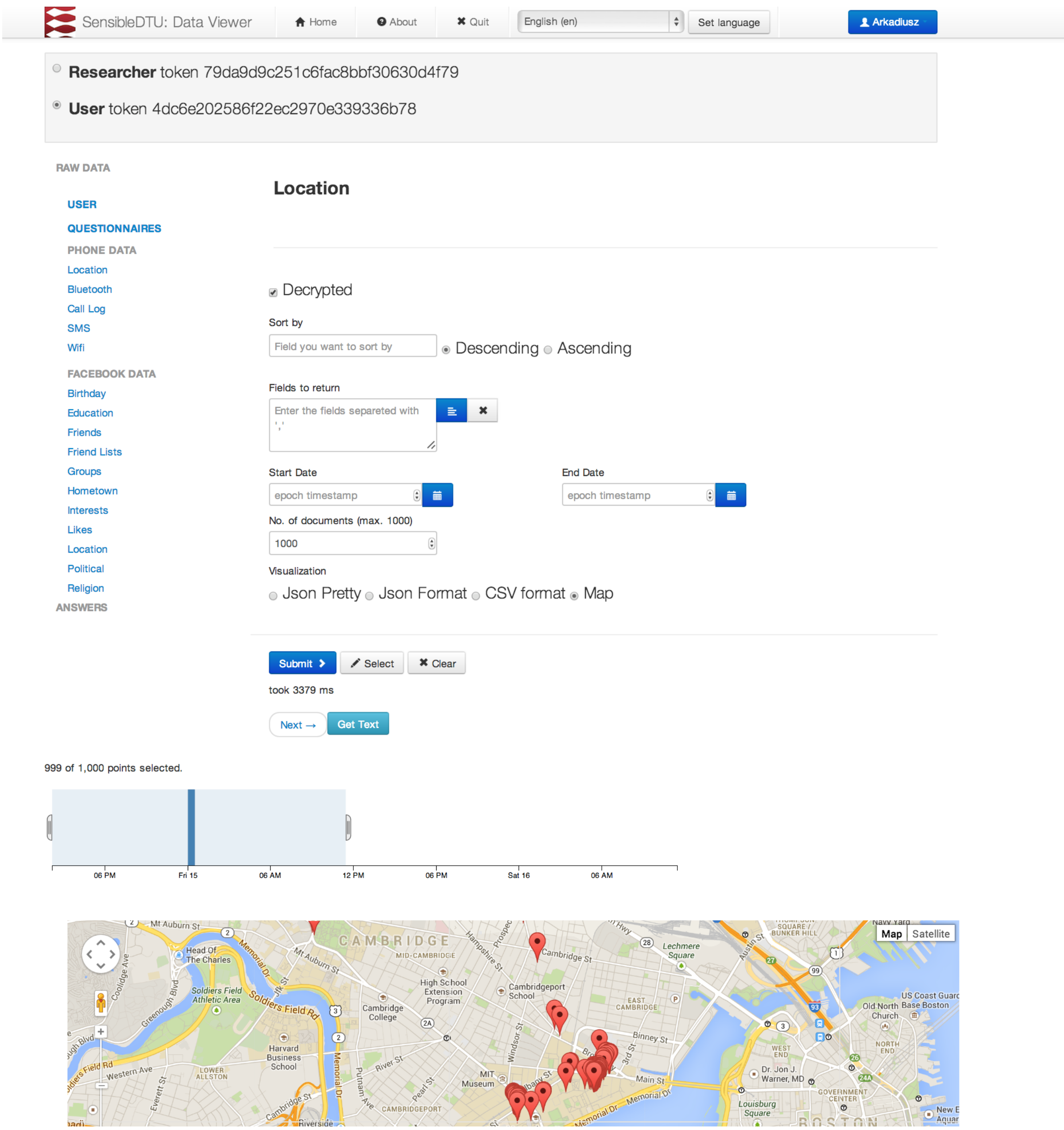


Figure 3. Data viewer application. All the collected data can be explored and access via an API. The API is the same for research, application, and end-user access, the endpoints are protected by OAuth2 bearer token.

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 users (set in 'visible' mode) can be used as a proxy for face-to-face interactions [28]. 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. Bluetooth scans do not constitute a perfect proxy for face-to-face interactions [142], 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 or to interact from a larger distance but, nevertheless,they have been successfully used for sensing social networks [30] or crowd tracking [143].

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 are triggered every five minutes, measured from the last time the phone was powered on. Thus, the phones scan for Bluetooth in a desynchronized fashion, 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 user i can observe user 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 user i observed user j, the opposite is also true.

The interactions between the participants exhibit both daily and weekly rhythms. Figure 4 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 cumulative degree distributions for varying temporal resolutions, with $P(k)$ being shifted towards higher degrees for larger window sizes; this is an expected behavior since each node has more time to amass connections. Figure 6b presents the opposite effect, where the edge weight distributions $P(w)$ shifts towards lower weights for larger windows, this is a consequence on definition of a link for longer time-scales or, conversely, of links appearing in each window on shorter timescales. To compare the distribution between timescales, we rescale the properties according to Krings et al. [144] as $P(x) \sim \langle x\rangle P(x/\langle x\rangle)$ with $\langle x\rangle = \sum xP(x)$ (Figure 6c and 6d). The divergence of the rescaled distributions suggest a difference in underlying social dynamics between long and short timescales, an observation supported by recent work on temporal networks [43, 144, 145].

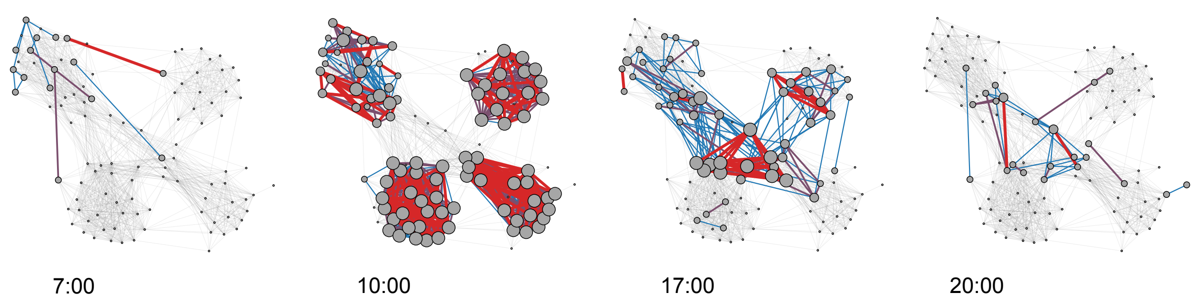


Figure 4. Dynamics of face-to-face interactions in SensibleDTU 2012. The participants meet in the morning, attend classes within four different study lines, and interact across majors in the evening. Edges are colored according to the frequency of observations, ranging from low (blue) to high (red). Node sizes are scaled according to degree.

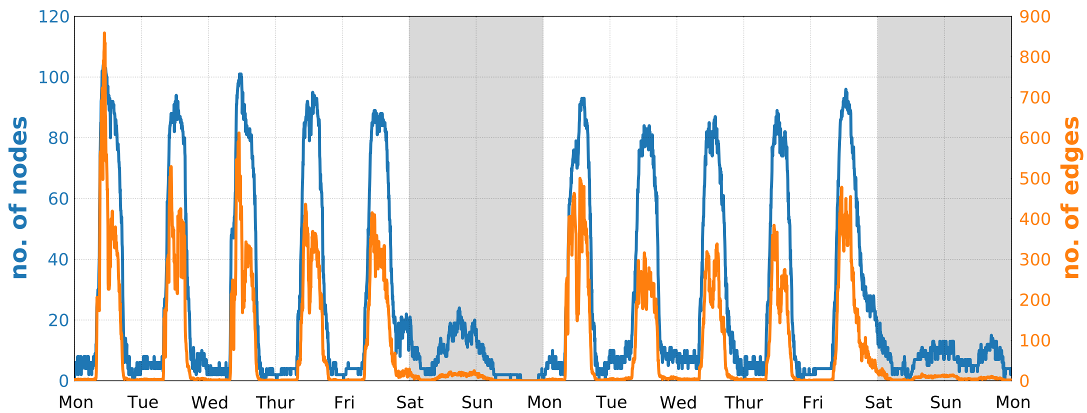


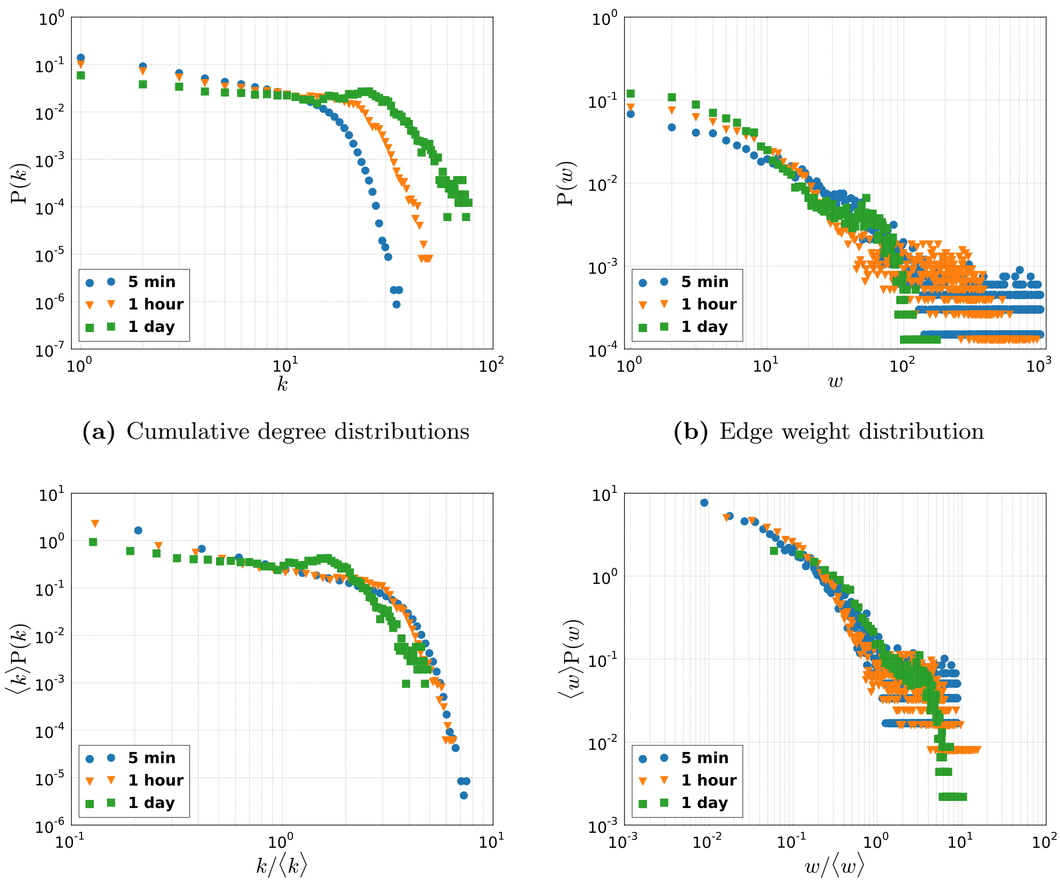
Figure 5. Weekly temporal dynamics of interactions. Face-to-face interaction patterns of participants in 5-minute time-bins over two weeks. Only active participants are included, i.e. users that have either observed another person or themselves been observed in a given time-bin. On average we observed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

7.2 WiFi as Additional Channel for Social Ties

For 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 [146] and wireless networks [147]. The SensibleDTU data collector application for mobile phones is 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 collect these data in an opportunistic way, 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 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 [18, 28, 30, 148]. The usage of WiFi as a social proxy has been investigated [149], but, to our knowledge, has not yet been used in a large-scale longitudinal study. For the method we describe here, the users' 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 which 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 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 signal 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 ($|X \cap Y|$)



(a) Cumulative degree distributions

(b) Edge weight distribution

(c) Rescaled degree

(d) Rescaled edge weight

Figure 6. Facet-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.

required to assume physical proximity. In approx. 98% of all Bluetooth encounters, at least one access point was seen by both devices, the recall however drops quickly with the increase of their required number. This measure favors interactions in places with 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 [150]. 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 $\mathrm{overlap}(X, Y) = \frac{|X \cap Y|}{\min(|X|, |Y|)}$ penalizes encounters taking place in WiFi-dense areas, due to higher probability of one device picking up a signal from a remote access point which is not available to the other device, see Figure 7b.

Next, we compare the received signal strengths between overlapping routers using the mean ℓ_1-norm (mean Manhattan distance, $\frac{||X \cap Y||_1}{|X \cap Y|}$). If all the overlapping AP's received signals with the same strength, the metric has value of zero; the bigger the differences in the signal strength, the higher the value. Mean Manhattan distance does not handle with situations in a desirable way. When devices see different access points, except for access seen with similar signal strength, this metric has a low value even though the observed access points are different.

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 interaction.

The performance of face-to-face event detection based on WiFi can be further improved by applying machine learning approaches [150, 151]. 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 there is a bias which 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.

7.3 Location and Mobility

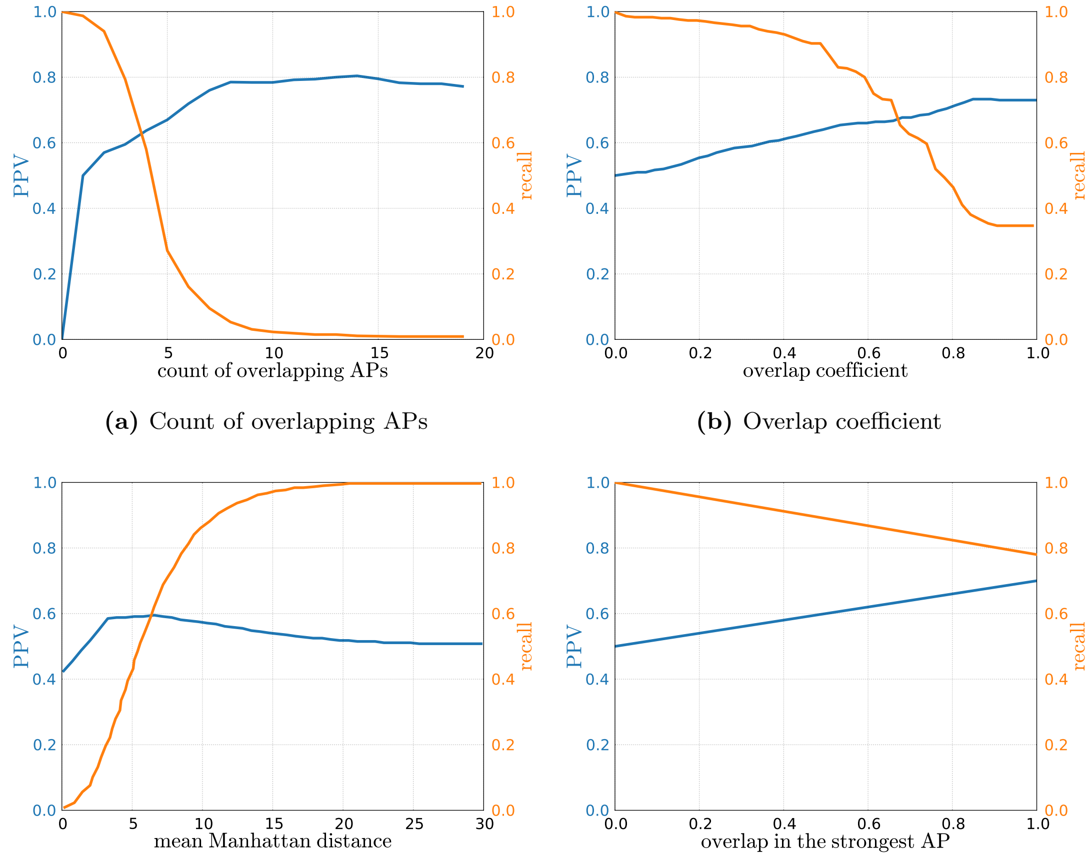
A number of applications ranging from urban planning, to traffic management, to containment of biological diseases rely on ability to accurately predict human mobility. Mining location data allows to extract semantic information such as points of interest, trajectories, and modes of transportation [152]. In this section we report preliminary results of an exploratory data analysis of location and mobility patterns in the SensibleDTU dataset.

Location data in SensibleDTU was obtained by periodically collecting the best position estimate from the location sensor on the 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 the form (userid, timestamp, latitude, longitude, accuracy). The best-effort nature of the data presents new challenges when compared to 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 `http://developer.android.com/guide/topics/location/strategies.html`. For this reason, accuracy of the collected position can vary between 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 reported accuracy better than 40 meters.

We calculate the radius of gyration r_g as defined in [37] and approximate the probability distribution function using a gaussian kernel density estimation, determining the bandwidth value by cross-validation (Figure 8b). The kernel density peaks around 10^2 km and then rapidly goes down, displaying a fat-tailed distribution. Manual inspection of the few users with r_g around 10^3 km revealed that travels abroad can amount to such high mobility. Although we acknowledge that this density estimation suffers due to the low number of samples, our measurements suggest that real user mobility is underestimated in studies based solely on CDRs, such as in [37], 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 [153–156]. By drawing edges between stop locations for each user, so that the most frequent transitions stand out, we can reveal patterns of collective mobility (Figure 8d).



(a) Count of overlapping APs

(b) Overlap coefficient

(c) Mean Manhattan distance

(d) Overlap of the strongest AP

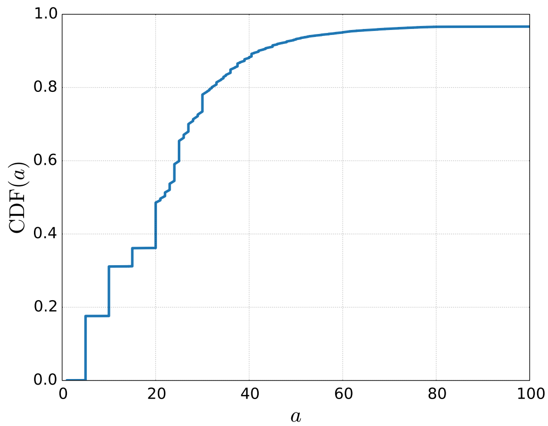
Figure 7. WiFi similarity measures. Positive predictive value (precision, ratio of number of true positives and number of positive calls) and recall (sensitivity, fraction of retrieved positives) as functions of parameters in different similarity measures. In 98% of face-to-face meetings derived from Bluetooth, the two devices have also sensed at least one common access point **(a)**. Identical strongest access point for two separate mobile devices, is a strong indication of a face-to-face meeting **(d)**.

7.4 Call & Text Communication Patterns

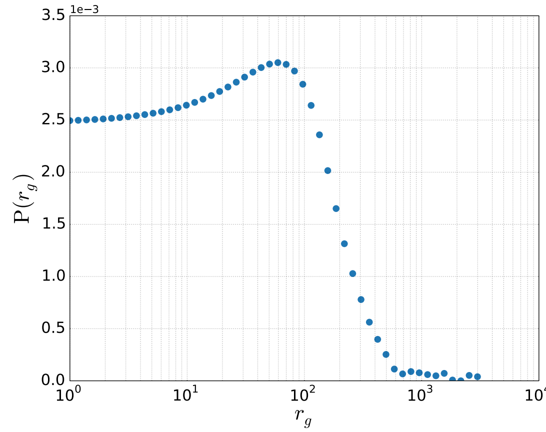
With the advent of mobile phones in the late 20$^{\text{th}}$ century, the way we communicate has changed dramatically. We are no longer restricted to landlines and are able to move around in physical space while communicating over long distances.

The ability to efficiently map communication networks and mobility patterns (using cell towers) for large populations, has made quantification of human mobility patterns possible, including investigations of social structure evolution [157], economic development [60], human mobility [36, 37], spreading patterns [55], and collective behavior with respect to emergencies [58]. In SensibleDTU 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).

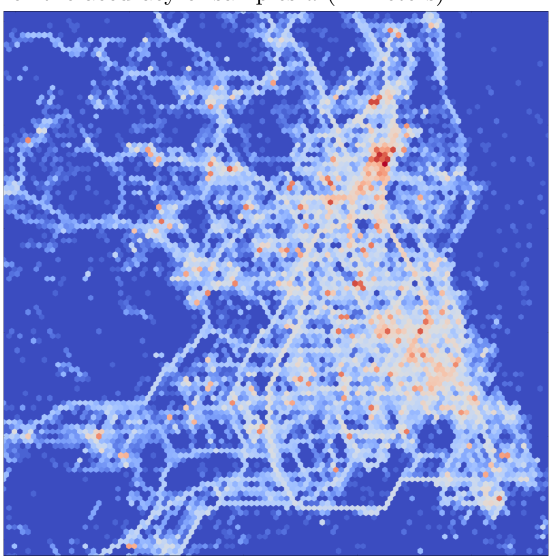
From October 1$^{\text{st}}$, 2012, to September 1$^{\text{st}}$, 2013, we collected 56 902 incoming and outgoing calls, of



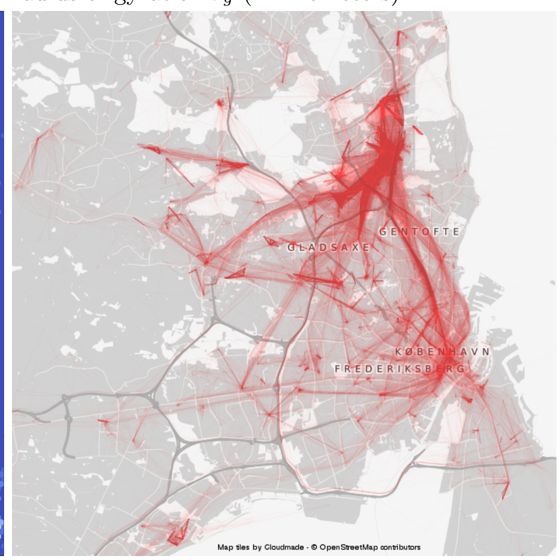
(a) Estimated cumulative distribution function for the accuracy of samples a (in meters).



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



(c) Two-dimensional histogram for 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 users, and identify patterns of collective mobility.

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

We find a correlation of 0.75 between the number of unique contacts users 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 users 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

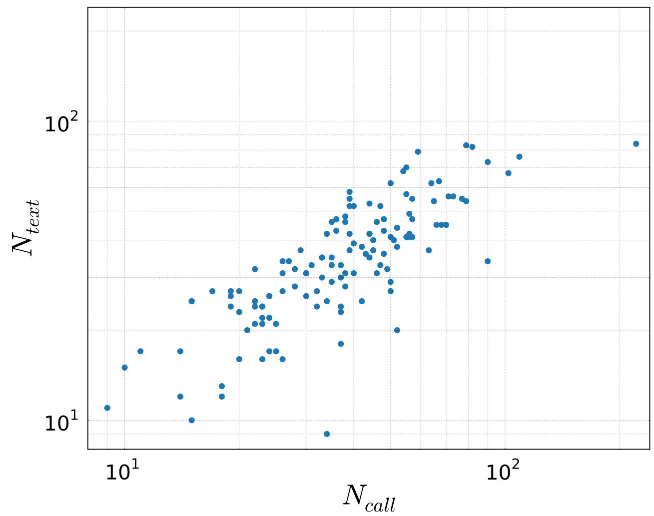


Figure 9. Diversity of communication logs. Diversity is estimated as the set of unique numbers that a person has contacted or been contacted by in the given time period on given channel. We note strong correlation in diversity, whereas the similarity of the sets of nodes is fairly low.

participants and with the external world) as a time series, calculated through the entire year and scaled to denote the mean count of interactions users had in given hourly time-bins in the week. Also here, we notice differences between those channels. While both clearly show decrease in activity during lunch time, call activity peaks around the end of business day and drops until next morning; SMS on the other hand, after similar decrease, which we can associate with commute, displays another evening peak. Also at night, SMS seems to be a more acceptable form of communication, with messages exchanges continuing late and starting early, especially on Friday night, when the party seems to never stop.

We point out that the call and SMS dynamics display patterns that are quite distinct from face-to-face interactions between participants showed in Figure 5. Although calls and SMS communication is 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 the university work, and only few of those ties manifest during the weekends, even as 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.

7.5 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 [158]. 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

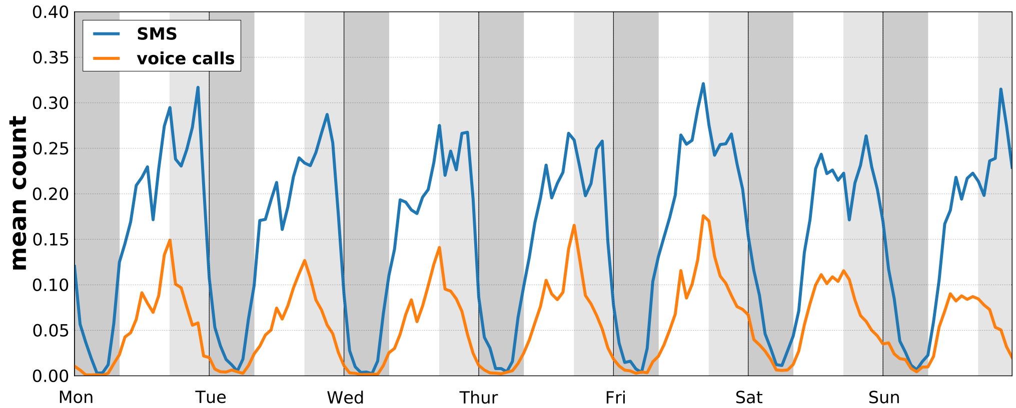


Figure 10. Weekly temporal dynamics of interactions. All calls and SMS', both incoming and outgoing calculated over the entire dataset and averaged per user and per week, showing mean number of interactions users had in given weekly bin. Light gray denotes 5pm, the end of lectures at the university,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

2013 deployment since, in contrast to the first deployment, we also collected interaction data. 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 disregarded. 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 which the face-to-face network misses—red edges in Figure 12.

7.6 Personality traits

While the data from mobile sensing and online social networks provide insights primarily into structure of social ties, we are also interested in the demographics, psychological and health traits, and interests of the participants. Knowing those characteristics, we can start answering questions about the reasons for the observed network formation; why are the ties created and what drives their dynamics? For example, homophily plays a vital role in how we establish, maintain, and destroy social ties [159].

Within SensibleDTU, participants answered questions covering the aforementioned domains. Those included a widely used *Big Five Taxonomy* [160], where five broad domains describing human personality are identified — openness, extraversion, neuroticism, agreeablenes, and conscientiousness. The traits are calculated from questionnaires by scoring questions from 1-5 (low to high) and calculating the average score of questions related to each personality domain.

As Big Five has been collected for various populations, including representative sample from Germany [161] and students from western Europe [162], we report the results from 2012 deployment in Figure 13 to show that our population is unbiased with respect to those important traits.

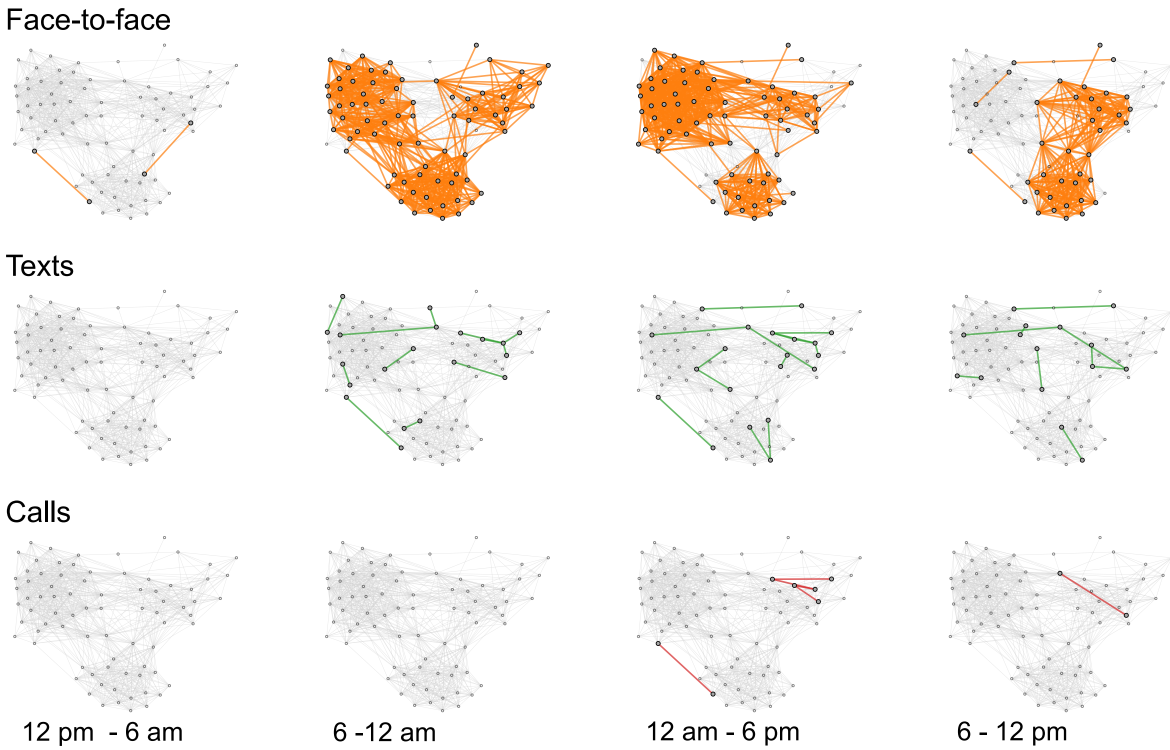


Figure 11. Daily activations in three networks. One day (Friday) in a network showing how different views are produced by observing different channels.

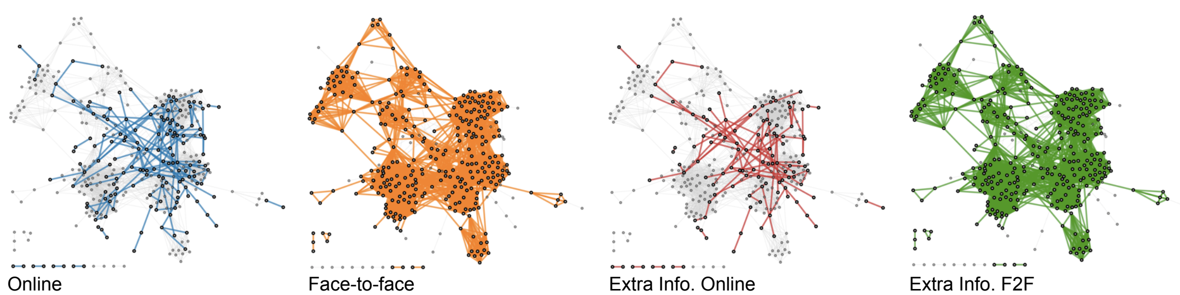


Figure 12. Face-to-face and online activity. Figure shows data from the 2013 deployment for one representative week. **Online:** Interactions (messages, wall posts, photos, etc.) between users on Facebook. **Face-to-Face:** Only the most active edges, which account for 80% of all traffic, are shown for 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.

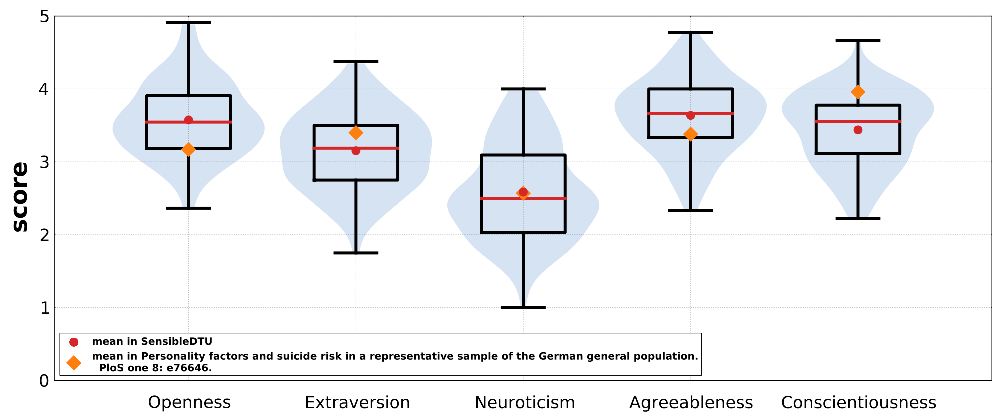


Figure 13. Personality traits Violin plot of personality traits. Summary statistics are: **openness** $\mu_O = 3.58$, $\sigma_O = 0.52$; **extraversion** $\mu_E = 3.15$, $\sigma_E = 0.53$; **neuroticism** $\mu_N = 2.59$ $\sigma_N = 0.65$; **agreeablenes** $\mu_A = 3.64$ $\sigma_A = 0.51$; **conscientiousness** $\mu_C = 3.44$ $\sigma_C = 0.51$. Mean values from SensibleDTU (red circles) compared with mean values reported in [161] (orange diamonds).

8 Perspectives

We expect that the amount of data collected about human beings will continue to increase. New and better services will be offered to the users, more effective advertising will be implemented, and researchers will be learning more about the human nature. As the complexity and scale of studies on social systems studies grows, collection of high-resolution data for studying human behavior will grow 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, the 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 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 related administrative effort of study management, may simply not be feasible.

In the not-so-distant future, many of the 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 users' authorizations, we can track their activity levels, using API's 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 replicability and reproducibility of the results—or selection bias in the context of full end-user data control. Still, we expect for the future studies to increasingly relay 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 SensibleDTU—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 stream of big data.

9 Acknowledgments

The SensibleDTU project was made possible by a Young Investigator Grant from the Villum Foundation (awarded to SL). Scaling the project up to 1000 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. We are indebted to our University of Copenhagen partners on a number of levels: All instrumentation on the 2013 questionnaires, as well as embedded anthropologist, are curtesy of the Social Fabric group, and most importantly the Social Fabric consortium has made valuable contributions through discussion and insight regarding nearly every aspect of the study. For an overview of the Social Fabric project, see `http://socialfabric.ku.dk/`. We thank group leaders in the Social Fabric Project: Professor David Dreyer Lassen, Professor Morten Axel Pedersen, Associate Professor Anders Blok, Assistant Professor Jesper Dammeyer, Associate Professor Joachim Mathiesen, Assistant Professor Julie Zahle, and Associate Professor Rikke Lund. From Institute of Economics: David Dreyer Lassen, Andreas Bjerre Nielsen, Anne Folke Larsen, and Nikolaj Harmon. From Institute of Psychology: Jesper Dammeyer, Lars Lundmann, Lasse Meinert Jensen, and Patrick Bender. From Institute of Anthropology: Morten Axel Pedersen, My Madsen. From Institute of Sociology: Anders Blok, Tobias Bornakke. From Institute of Philosophy: Julie Zahle. From 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

1. 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.

2. Aral S, Walker D (2012) Identifying influential and susceptible members of social networks. Science 337: 337–341.

3. Pietri R (2013). Privacy in computational social science. URL `http://www.compute.dtu.dk/English.aspx`. DTU supervisor: Sune Lehmann Jørgensen, sljo@dtu.dk, DTU Compute.

4. Onnela JP, Saramäki J, Hyvönen J, Szabo G, Lazer D, et al. (2007) Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences, USA 104: 7332.

5. Cha M, Haddadi H, Benevenuto F, Gummadi PK (2010) Measuring user influence in twitter: The million follower fallacy. ICWSM 10: 10–17.

6. Shalizi CR, Thomas AC (2011) Homophily and contagion are generically confounded in observational social network studies. Sociological Methods & Research 40: 211–239.

7. Barabâsi 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.

8. Kossinets G, Watts DJ (2006) Empirical analysis of an evolving social network. Science 311: 88–90.

9. Lazer D, Pentland A, Adamic L, Aral S, Barabasi A, et al. (2009) Life in the network: the coming age of computational social science. Science (New York, NY) 323: 721.

10. Wesolowski 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.

11. Madrigal A (2013) Dark social: We have the whole history of the web wrong. The Atlantic .

12. 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.

13. 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 .

14. Raento M, Oulasvirta A, Eagle N (2009) Smartphones an emerging tool for social scientists. Sociological methods & research 37: 426–454.

15. Chronis I, Madan A, Pentland AS (2009) Socialcircuits: the art of using mobile phones for modeling personal interactions. In: Proceedings of the ICMI-MLMI'09 Workshop on Multimodal Sensor-Based Systems and Mobile Phones for Social Computing. ACM, p. 1.

16. Pentland AS (2008) Honest signals: how they shape our world. MIT Press.

17. Olguín 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.

18. Miller G (2012) The smartphone psychology manifesto. Perspectives on Psychological Science 7: 221–237.

19. Raento M, Oulasvirta A, Petit R, Toivonen H (2005) Contextphone: A prototyping platform for context-aware mobile applications. Pervasive Computing, IEEE 4: 51–59.

20. 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.

21. 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.

22. 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.

23. 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.

24. Kostakos V, ONeill 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.

25. 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.

26. Hu X, Chu TH, Chan HC, Leung VC (2013) Vita: A crowdsensing-oriented mobile cyber physical system .

27. Larsen JE, Jensen K (2009) Mobile context toolbox. In: Smart Sensing and Context, Springer. pp. 193–206.

28. Eagle N, Pentland A (2006) Reality mining: sensing complex social systems. Personal and Ubiquitous Computing 10: 255–268.

29. (2013). "funf open sensing framework". URL `http://funf.org/`.

30. 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 .

31. Kiukkonen N, Blom J, Dousse O, Gatica-Perez D, Laurila J (2010) Towards rich mobile phone datasets: Lausanne data collection campaign. Proc ICPS, Berlin .

32. 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.

33. Olguín D, Waber 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.

34. Karikoski J, Nelimarkka M (2011) Measuring social relations with multiple datasets. International Journal of Social Computing and Cyber-Physical Systems 1: 98–113.

35. 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.

36. Gonzalez MC, Hidalgo CA, Barabasi AL (2008) Understanding individual human mobility patterns. Nature 453: 779–782.

37. Song C, Qu Z, Blumm N, Barabási AL (2010) Limits of predictability in human mobility. Science 327: 1018–1021.

38. 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.

39. Bagrow JP, Lin YR (2012) Mesoscopic structure and social aspects of human mobility. PloS one 7: e37676.

40. De Domenico M, Lima A, Musolesi M (2012) Interdependence and predictability of human mobility and social interactions. arXiv preprint arXiv:12102376 .

41. 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.

42. Eagle N, Pentland AS (2009) Eigenbehaviors: Identifying structure in routine. Behavioral Ecology and Sociobiology 63: 1057–1066.

43. Clauset A, Eagle N (2012) Persistence and periodicity in a dynamic proximity network. arXiv preprint arXiv:12117343 .

44. Onnela JP, Saramäki J, Hyvönen J, Szabó G, De Menezes MA, et al. (2007) Analysis of a large-scale weighted network of one-to-one human communication. New Journal of Physics 9: 179.

45. Granovetter MS (1973) The strength of weak ties. American journal of sociology : 1360–1380.

46. 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.

47. Onnela JP, Arbesman S, González MC, Barabási AL, Christakis NA (2011) Geographic constraints on social network groups. PLoS one 6: e16939.

48. Hidalgo CA, Rodriguez-Sickert C (2008) The dynamics of a mobile phone network. Physica A: Statistical Mechanics and its Applications 387: 3017–3024.

49. Miritello G, Lara R, Cebrian M, Moro E (2013) Limited communication capacity unveils strategies for human interaction. Scientific reports 3.

50. 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 .

51. 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.

52. 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.

53. Madan A, Cebrian M, Moturu S, Farrahi K, Pentland S (2012) Sensing the health stateof a community .

54. Madan A, Farrahi K, Gatica-Perez D, Pentland A (2011) Pervasive sensing to model political opinions in face-to-face networks. Pervasive Computing : 214–231.

55. Wang P, González MC, Hidalgo CA, Barabási AL (2009) Understanding the spreading patterns of mobile phone viruses. Science 324: 1071–1076.

56. 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.

57. 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.

58. Bagrow JP, Wang D, Barabási AL (2011) Collective response of human populations to large-scale emergencies. PLoS one 6: e17680.

59. Karsai M, Perra N, Vespignani A (2013) The emergence and role of strong ties in time-varying communication networks. arXiv preprint arXiv:13035966 .

60. Eagle N, Macy M, Claxton R (2010) Network diversity and economic development. Science 328: 1029–1031.

61. Blondel V, Krings G, Thomas I (2010) Regions and borders of mobile telephony in belgium and in the brussels metropolitan zone. Brussels Studies 42.

62. 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.

63. Klasnja P, Consolvo S, Choudhury T, Beckwith R, Hightower J (2009) Exploring privacy concerns about personal sensing. Pervasive Computing : 176–183.

64. 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.

65. 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.

66. 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.

67. Srivatsa M, Hicks M (2012) Deanonymizing 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.

68. 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.

69. 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.

70. 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.

71. 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.

72. 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.

73. Sweeney L (2000) Simple demographics often identify people uniquely. Health (San Francisco) : 1–34.

74. Barbaro M, Zeller T, Hansell S (2006) A face is exposed for aol searcher no. 4417749. New York Times 9: 8For.

75. Sweeney L (2002) k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10: 557–570.

76. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M (2007) l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD) 1: 3.

77. 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.

78. 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.

79. Dwork C, Nissim K (2004) Privacy-preserving datamining on vertically partitioned databases. In: Advances in Cryptology–CRYPTO 2004. Springer, pp. 134–138.

80. 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.

81. 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.

82. Chawla S, Dwork C, McSherry F, Smith A, Wee H (2005) Toward privacy in public databases. Theory of Cryptography : 363–385.

83. Rivest RL, Adleman L, Dertouzos ML (1978) On data banks and privacy homomorphisms. Foundations of secure computation 4: 169–180.

84. Gentry C (2009) A fully homomorphic encryption scheme. Ph.D. thesis, Stanford University.

85. TEBAA M, EL HAJJI S, EL GHAZI A (2012) Homomorphic encryption applied to the cloud computing security. Lecture Notes in Engineering and Computer Science 2197.

86. Naehrig 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.

87. 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.

88. 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.

89. Zdancewic SA (2002) Programming languages for information security. Ph.D. thesis, Cornell University.

90. Sfaxi L, Abdellatif T, Robbana R, Lakhnech Y (2010) Information flow control of component-based distributed systems. Concurrency and Computation: Practice and Experience .

91. Zeldovich N, Boyd-Wickizer S, Mazieres D (2008) Securing distributed systems with information flow control. NSDI.

92. Mundada Y, Ramachandran A, Feamster N (2011) Silverline: Data and network isolation for cloud services. Proc of 3rd HotCloud .

93. Pappas V, Kemerlis V, Zavou A, Polychronakis M, Keromytis AD (2012) Cloudfence: Enabling users to audit the use of their cloud-resident data .

94. Ganjali A, Lie D (2012) Auditing cloud administrators using information flow tracking .

95. Boneh D, Lipton R (1996) A revocable backup system. In: USENIX Security Symposium. pp. 91–96.

96. Perlman R (2005) The ephemerizer: Making data disappear .

97. Perlman R (2005) File system design with assured delete. In: Security in Storage Workshop, 2005. SISW'05. Third IEEE International. IEEE, pp. 6–pp.

98. 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.

99. Agrawal R, Haas PJ, Kiernan J (2003) Watermarking relational data: framework, algorithms and analysis. The VLDB journal 12: 157–169.

100. 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.

101. Cox IJ, Linnartz JP (1998) Some general methods for tampering with watermarks. Selected Areas in Communications, IEEE Journal on 16: 587–593.

102. 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.

103. 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`.

104. 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.

105. 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.

106. Mucha P, Richardson T, Macon K, Porter M, Onnela JP (2010) Community Structure in Time-Dependent, Multiscale, and Multiplex Networks. Science 328: 876-878.

107. 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.

108. Madden M, Lenhart A, Cortesi S, Gasser U, Duggan M, et al. (2013). Teens, Social Media, and Privacy. `http://www.pewinternet.org/~/media//Files/Reports/2013/PIP_TeensSocialMediaandPrivacy.pdf`.

109. Kossinets G (2006) Effects of missing data in social networks. Social Networks 28: 247–268.

110. Laumann E, Marsden P, Prensky D (1983) The boundary specification problem in network analysis, London: Sage Publications. pp. 18–34.

111. Saramki J, Leicht EA, Lpez E, Roberts SGB, Reed-Tsochas F, et al. (2014) Persistence of social signatures in human communication. Proceedings of the National Academy of Sciences .

112. Holme P, Saramäki J (2011). Temporal network. arxiv/1108.1780.

113. Lyons R (2011) The spread of Evidence-Poor medicine via flawed Social-Network analysis. Statistics, Politics, and Policy 2.

114. Shalizi C, Thomas A (2011) Homophily and contagion are generically confounded in observational social network studies. Sociological Methods & Research 40: 211–239.

115. 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.

116. Christakis N, Fowler J (2009) Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives. Little, Brown and Company.

117. 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.

118. Larsen JE, Cuttone A, Jørgensen SL Qs spiral: Visualizing periodic quantified self data. In: CHI 2013 Workshop on Personal Informatics in the Wild: Hacking Habits for Health & Happiness.

119. 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.

120. Rocha L, Liljeros F, Holme P (2011) Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. PLoS Computational Biolology 7: e1001109.

121. Lee S, Rocha LC, Liljeros F, Holme P (2010). Exploiting temporal network structures of human interaction to effectively immunize populations. arXiv/1011.3928.

122. Fortunato S (2010) Community detection in graphs. Physics Reports 486: 75 - 174.

123. Gulbahce N, Lehmann S (2008) The art of community detection. BioEssays 30: 934–938.

124. Ahn YY, Bagrow JP, Lehmann S (2010) Link communities reveal multiscale complexity in networks. Nature 466: 761-764.

125. Fiedler M (1975) A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. Czechoslovak Mathematical Journal 25: 619.

126. Bagrow JP, Lehmann S, Ahn YY (2011). Robustness and modular structure in networks. arxiv/1102.5085.

127. (2013). Facebook reports first quarter 2013 results. URL `http://investor.fb.com/releasedetail.cfm?ReleaseID=761090`.

128. Costa PT, MacCrae RR (1992) Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO FFI): Professional Manual. Psychological Assessment Resources.

129. Rosenberg M (1989) Society and the adolescent self-image (rev. Wesleyan University Press.

130. Back MD, Küfner AC, Dufner M, Gerlach TM, Rauthmann JF, et al. (2013) Narcissistic admiration and rivalry: Disentangling the bright and dark sides of narcissism. .

131. Diener E, Emmons RA, Larsen RJ, Griffin S (1985) The satisfaction with life scale. Journal of personality assessment 49: 71–75.

132. Rotter JB (1966) Generalized expectancies for internal versus external control of reinforcement. Psychological monographs: General and applied 80: 1.

133. Russell DW (1996) Ucla loneliness scale (version 3): Reliability, validity, and factor structure. Journal of personality assessment 66: 20–40.

134. 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.

135. Cohen S, Kamarck T, Mermelstein R (1983) A global measure of perceived stress. Journal of health and social behavior : 385–396.

136. 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.

137. 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.

138. 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.

139. Shampanier K, Mazar N, Ariely D (2007) Zero as a special price: The true value of free products. Marketing Science 26: 742–757.

140. (2013). SensibleDTU informed consent form (da). URL `https://github.com/MIT-Model-Open-Data-and-Identity-System/SensibleData-Service/blob/production_sensibledtu1k/sensible_data_service/documents/service_informed_consent_da.txt`.

141. (2013). SensibleDTU informed consent form (en). URL `https://github.com/MIT-Model-Open-Data-and-Identity-System/SensibleData-Service/blob/production_sensibledtu1k/sensible_data_service/documents/service_informed_consent_en.txt`.

142. 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 .

143. Stopczynski A, Larsen JE, Lehmann S, Dynowski L, Fuentes M Participatory bluetooth sensing: A method for acquiring spatio-temporal data about participant mobility and interactions at large scale events .

144. 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.

145. Ribeiro B, Nicola P, Baronchelli A (2013) Quantifying the effect of temporal resolution on time-varying networks. Scientific reports 3.

146. Whitehead M, Phillips T, Page M, Molina M, Wood C (2012). European mobile industry observatory 2011. `http://www.gsma.com/publicpolicy/wp-content/uploads/2012/04/emofullwebfinal.pdf`.

147. 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.

148. Madan A, Cebrian M, Moturu S, Farrahi K, Pentland S (2011) Sensing the health stateof a community. Pervasive Computing .

149. Kjærgaard 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.

150. 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.

151. Carreras I, Matic A, Saar P, Osmani V (2012) Comm2Sense: Detecting Proximity Through Smartphones. PerMoby Workshop, part of IEEE PerCom 12 Conference .

152. Lin M, Hsu WJ (2013) Mining gps data for mobility patterns: A survey. Pervasive and Mobile Computing .

153. Hariharan R, Toyama K (2004) Project lachesis: parsing and modeling location histories. In: Geographic Information Science, Springer. pp. 106–124.

154. 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.

155. 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.

156. 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.

157. Palla G, Barabási AL, Vicsek T (2007) Quantifying social group evolution. Nature 446: 664–667.

158. (2013). "Facebook Reports Third Quarter 2013 Results". URL http://investor.fb.com/releasedetail.cfm?ReleaseID=802760.

159. McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: Homophily in social networks. Annual review of sociology : 415–444.

160. John OP, Naumann LP, Soto CJ (2008) Paradigm shift to the integrative big five trait taxonomy. Handbook of personality: Theory and research 3: 114–158.

161. 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.

162. 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.