Measuring Patterns of Human Behavior through Fixed-Location Sensors

Piotr Sapieżyński

Supervisors:
Sune Lehmann
Jakob Eg Larsen

Kongens Lyngby 2013
IMM-M.Sc.-2013-18

Technical University of Denmark
DTU Compute — Department of Applied Mathematics and Computer Science
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk IMM-M.Sc.-2013-18

Summary (English)

Fixed-location sensor networks are growing in popularity as they provide a cost-effective solution for measuring behavior of large populations. In many cases, existing infrastructure can be turned into a sensor network by using custom software; alternatively, off-the-shelf hardware can be used to build such networks from scratch. Since collecting data through such sensors does not require any interaction with the observation subjects, it does not influence their behavior, and can thus provide objective insights. This very same characteristic makes fixed-location sensing controversial in terms of privacy protection, as it can be deployed without any consent from the observed persons.

This thesis provides a general review of the fixed-location sensor networks in use currently, and the state of the art statistical approaches to modeling the collected data. It discusses possible directions of future development of fixed-location sensing as well as privacy and security implications of such approach to behavioral tracking. Finally, analysis of real world data collected during experimental deployments in different contexts and the fixed-location sensor perspective is compared to mobile sensor perspective, to show the following:

- It is possible to capture mobility and interaction traces of crowd participants through stationary Bluetooth scanners; modeling such data using Infinite Relational Model reveals underlying motivations of the participants.
- Stationary WiFi Access Points provide a highly accurate insight into social networks based on physical proximity (who *spends time* with whom); it is possible to extract behavioral patterns from such data and, through

mathematical modeling, recover the cognitive social network (who *likes* whom).

Additionally, the thesis presents one paper, which I authored with researchers from Alan Mislove's Theory and Networked Systems Lab and David Lazer's LazerLab at Northeastern University, which however falls outside of the main scope of the thesis.

Preface

This thesis was prepared at DTU Compute - Department of Applied Mathematics and Computer Science at the Technical University of Denmark in fulfillment of the requirements for acquiring a M.Sc. in Digital Media Engineering.

The thesis deals with distributed, fixed-location sensing of human presence and co-presence, as well as mathematical approaches employed to derive higher-level information such as cognitive social networks.

It consists of three articles that scrutinize different approaches to sensing and mathematical modeling, each based on real-world data.

Lyngby, 15-March-2013

Piotr Sapieżyński

Acknowledgements

First and foremost, I would like to thank my tutor, Tobias Andersen, for his support throughout my studies in Denmark, ever since I took his course as an Erasmus student with no intention of staying at DTU, all the way to where I am now.

I also wish to thank Sune Lehmann, whose course ignited my love for those colorful network visualizations, which you will find in the thesis - an interest, which gradually turned into a slightly more serious desire to become a scientist. I am also extremely grateful to Sune for helping me make the first steps towards fulfilling this desire, as well as for facilitating my research stay in Boston; the six months in USA were the time of my life, and not only academically.

I thank Morten Mørup for the fruitful discussions at the Chinese Restaurant and for his irreplaceable guidance through the Random Forest.

I wish to express my gratitude to David Lazer[1], for giving me an opportunity to work in the World's hub of Computational Social Science, with the very scientists whose work inspired me to pursue an academic career.

None of the projects described in the thesis would have been possible without the hacking skills of Arek Stopczyński and his ability to make me feel constantly bad about not working enough. Thank you, Arek.

Special thanks go to Søren Kamaric Riis, Steen Munk, Anders Meng, and my other colleagues from Oticon for creating a wonderful work place atmosphere,

[1] pronounced as if it were Lazar

helping me develop my skills, and supporting my decisions.

I would like to thank all the people, who I did not thank yet, and with whom I had the pleasure to work on the projects described in this thesis: Jakob Eg Larsen, Aniko Hannak, Arash Molavi Kakhki, Alan Mislove, Christo Wilson, Balachander Krishnamurthy, Rasmus Theodorsen, and Vedran Sekara.

Many thanks to my tireless proof-readers, whose efforts I hope not to have put to waste during the last-minute general overhaul of the thesis: Anna Tatarczak, Sven MacAller, Gabriel Aldaz, and Piotr Borowian.

Finally, I thank my Parents, for their never-ending patience towards my never-ending studies.

Contents

Chapter 1

Introduction

Observation and qualitative analysis of behavior and interactions between humans gave ground to such sciences as anthropology, psychology, and sociology. While they have formed our today's understanding of individuals, groups, and societies, they have invariably suffered from limited scope, or biases introduced by participant observation, or both, as longitudinal and thorough examination of a group of people used to be impossible to perform through unobtrusive measures. This situation has however been changing, due to technological and computational advancements.

Over the last few years, mobile phones have become increasingly ubiquitous, not only revolutionizing communications but also providing a previously unthinkable framework for gaining insights into the structure of societies. From mobility of a single user, to communication patterns within small communities, to complex models of dynamic social networks spanning across the world – all these topics which were very hard, or even unfeasible to study just a few decades ago, are now being explored with the help of the records of mobile network providers. With the advent of smartphones, which constantly send and receive data without the user's active participation, these logs are becoming even more representative, as they now provide the information about the users with a very high temporal resolution, not only on rare instances of making a phone call, or sending a message.

However, the increasing availability of longitudinal, spatio-temporal traces of human lives is not a by-product of just the mobile phones adaptation. In the computerized world, virtually all actions people take leave some kind of *electronic breadcrumbs*, which are used to build a comprehensive history of each person's behavior. Whenever a person uses a credit card to pay for shopping, the card provider and the bank notes *who* bought *what*, *how much* they paid and *where* they did the shopping. Whenever a person browses the Web, countless companies take notice of every single visited page, see Figure 1.1. Public transportation companies provide "free" access to WiFi, but in return they learn *who* travels *from where*, *to where*, *how long* these trips take and *what time of day* they occur. There are many other examples of situations, where people leave collectable traces of their actions, and listing them, while interesting and eye-opening, is outside of scope of this thesis.

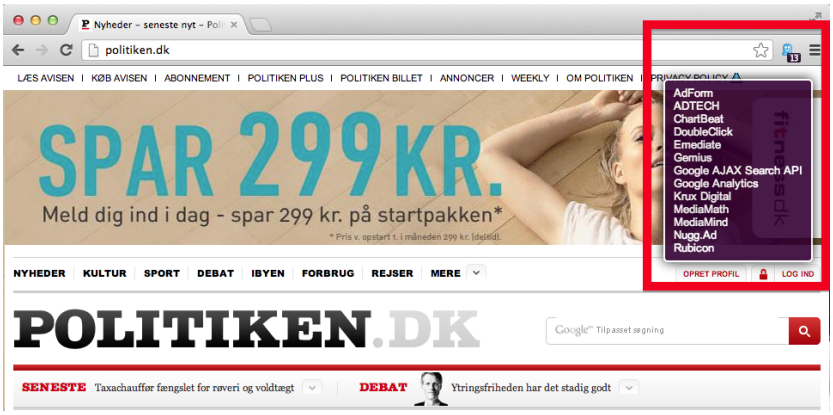


Figure 1.1: A screenshot of Google Chrome browser with Ghostery[1] plugin, which lists the tracking cookies and prevents them from being deployed into the user's machine. Visiting the newspaper Politiken's website is recorded by 12 different companies. Interestingly, the usually ubiquitous Facebook and Twitter cookies are not present on this webpage

In all the situations listed above, some service is offered to a potential user without requiring them to pay with money (at least not the full monetary price of the service), and the user "pays" by revealing some information about them. While this transaction does not require any consent, an aware person does expect that it takes place. On the other hand, there are scenarios, which do not require any action from the observed persons, other than merely to be in a particular place, at a particular time, such as video surveillance, or Bluetooth scanning. Such sources of information provide an even more objective insight into human

[1] http://www.ghostery.com/

behavior, but are also much more controversial.

In this thesis, I will start by describing the findings of researchers, who applied some of the above-mentioned methods, which I will refer to as *fixed-location sensors*, for collecting behavioral data, and compare these insights to those gained by using *mobile sensors*. I will also give an outline on measures taken to protect the privacy of the observed participants. Finally, I will present three examples of experimental deployments of sociometric studies, which utilize fixed-location sensors. The first of the presented studies is Bluetooth scanning at Roskilde music festival, described in Chapter 2. I show that this method of scanning is suitable for collecting mobility and interaction traces of thousands of people using cheap, off-the shelf hardware, and that modeling of the collected data can reveal underlying motivations of the Festival's participants. The second study examines the quality of data collected using fixed location WiFi Access Points and GSM towers by comparing it to the ground truth data obtained by mobile Bluetooth scanning, described in Chapter 3. I show that Call Detail Records, which are widely used in Computational Social Science to study mobility, cannot be used to infer physical proximity with satisfying results, while WiFi access points are much more suitable for this task. The third example deals with recovering cognitive network by extracting high-level features from behavioral traces as recorded by WiFi APs, described in Chapter 4. It shows that state of the art machine learning approaches can be used on such data to recover real-world friendship networks. In the Appendix, you will find a paper, which deals with another approach to user-modeling: the personalization of Web search. Note, there is some repetition across the chapters, as each of them is a separate paper and must be understood without reading this introduction, or other chapters.

1.1 Computational social science

The term *computational social science* was coined by Lazer et al. [LPA+09] to mark the transition from traditional, survey-based social science, to the modern, data-driven approach. As described by Lazer, computational social scientists take advantage of the emerging *sociometer* tools, such as radio frequency identification (*RFID*) badges, video surveillance, email analysis, and smartphones to collect continuous behavioral data with high resolution, as opposed to traditional social scientists, who would based their research on one-off surveys and participative observation. The computational approach to analyzing human behavioral patterns was not started in academia, but in Web companies, like Google and AOL, mobile network providers, and intelligence agencies.

1.2 Methods and challenges in social sensing

Mobile network providers were the first to analyze human mobility based on Call Detail Records (CDR), currently a very popular approach to computational analysis of societies. Initially, predictive models of mobility of individuals were developed in order to aid the development of network infrastructure and hand-off procedures [LM95]. Only recently have social scientists started investigating the CDR data. In a study from 2008, Gonzáles et al. [GHB08] quantified the predictability of users' location based on the CDRs of 100,000 people. They showed that distributions of basic mobility features, like displacement between samples, and radius of gyration, can be approximated by the power law, and the travel patterns of individuals can be approximated by a Lévy flight[2]. Gonzáles et al. concentrated on individual users, without investigating the possible interaction between them, and its influence on their mobility. Wang et al. [WPS+11] used CDRs from 6 million mobile phone users to show, that while colocation inferred from cell towers connectivity is correlated with social ties, link prediction based on the data yields low precision and recall. The difficulty in using this kind of data is caused by its inherently irregular sampling. Call records are only created whenever a phone call is placed, or an SMS exchanged. The sampling problem is described in great detail in [RZZB12], by comparing the traditional CDRs to their mobile data records counterpart. Since much of the data traffic on smartphones is periodical check for updates, the data records have much more stable and representative sampling than call based records. In Chapter 3 I describe our attempt at leveraging the call records for reconstructing face-to-face meetings.

Various studies analyzed data from WiFi networks to model the mobility of their users, but most of them focused on analyzing single-user patterns, while ignoring the social aspect. For example, Ravi et al. [JLB05] used real-world traces of 6000+ students collected over 2 years, to develop a model, which then they used for simulating the network load (note, that the mobility of individuals is again used for quality of service assurance, rather than investigating humans themselves). They showed that the mobility patterns are far from random, and therefore the random walk model does not reflect the reality of a campus network. In another study, Griswold et. al [GBB+02] leveraged proximity sensed using the system WiFi to provide social functionality to users during the Active Campus project. However, traces collected from 700 mobile devices handed out to participants had never been used for studying social interactions. Using system perspective WiFi[3], might yield more accurate results than using CDRs

[2] Lévy flight is a random walk with step length characterized by a heavy-tailed probability distribution

[3] system perspective - seen from the perspective of the network provider, thus limited only to the AP a user is connected to, as opposed to a much richer user perspective, where the

in terms of spatial resolution and sampling rate, but also creates a number of challenges which do not occur with mobile network data. Firstly, while mobile networks cover majority of areas of developed countries (98% of geographic coverage in Europe in 2011 according to [WPP+12]), the WiFi networks are, at best, citywide, and usually much smaller. Furthermore, while virtually everybody uses a mobile phone and can be tracked using the CRDs (100% population coverage according to [WPP+12]), a significantly smaller part uses mobile devices featuring WiFi, and even those do not connect to all available WiFi networks. Still, given all these difficulties, the system perspective WiFi signal is a very valuable source of information about the human mobility and interactions, and can be used very effectively to reconstruct real-world social networks, at least in specific contexts, as I show in Chapter 4.

The above-mentioned approaches, CDR and system WiFi data analysis, leverage existing physical infrastructure and embedded data collection mechanisms to obtain the data, and then use purpose-built software to extract information meaningful to computational social scientists. The elementary advantage of such methods is that no additional infrastructural work needs to be done, thus significantly reducing the cost of deploying experiments. Moreover, these existing networks provide very valuable service to people and therefore, by being extensively utilized, they are able to collect vast amount of real-world data. However, as pointed out before, the drawback is that they are not designed with behavioral observations in mind, and are thus not perfectly suited for such application.

Purpose-built sensor networks are an interesting alternative, as they can be constructed using off-the-shelf hardware and deployed in a way, which fulfills the observation/experiment needs. Networks of Bluetooth scanners are one example of such an approach. The hardware they utilize is cheap and readily available, as even the least expensive smartphones or computers can be used as sensors. Bluetooth is by design a technology for exchanging data over short distances (up to 10 meters in consumer devices), so the scanners offer far better spatial resolution than the GSM masts or WiFi APs can provide. Short-term observations can be deployed in arbitrary locations as the smartphone-based scanners can run several hours on a single battery load, but longer lasting deployments are limited to locations with power supply. A number of research groups have been using Bluetooth scanners for measuring mobility and sensing interactions of people. Most notably, O'Neill, Kostakos, et al. [OKK+06, KOP+10] deployed a number of such devices in semantically varied locations around the city. They analyzed the behavioral traces of visitors of these places, showing profound differences in patterns of presence and co-presence in locations with

participant's device can sense all the nearby APs along with the received signal strength indication, without a need to connect to any of them

different social function, for example a busy street vs. a bar. Unfortunately, even though their study involved deploying more than 90 scanners, only four specific locations are described in the papers - a street, the university entrance, an office and a bar. It is not clear how much insight O'Neill et al. gained into the social structure of other locations. While Bluetooth sensor networks have high spatial and virtually arbitrary time resolutions, and are cheap to deploy, they are gradually becoming less popular. The main drawback of social sensing through Bluetooth is that the number of people "visible" over this channel is declining with the growing penetration of smartphones. All Android [And13], iOS [App13], BlackBerry [Bla13] and WindowsPhone [Win13] devices ship with discoverability turned off. Combined smartphone market share of these devices was estimated to 96.9% in 4Q12 by IDC [IDC13], while the smartphone market penetration in Denmark was estimated to 45% in 1Q12 by Google [Goo12]. Moreover, iPhones and Windows Phone devices cannot be permanently discoverable, rendering them invisible to Bluetooth scanners. Even though Bluetooth scanning cannot be expected to provide comprehensive information about each individual in range, it can still be used to sense general trends in large groups of people. In Chapter 2 I describe an example of a Bluetooth scanners sensor network deployment and mathematical models, which can be applied to extract meaningful data from the raw logs.

1.3 Modeling

Regardless of which physical method is chosen for behavioral sensing, the raw collected data is not informative without processing and modeling. In this section I present an overview of mathematical approaches currently being employed in analysis of the mobility of individuals, peer influence on their behavioral patterns, as well as in network modeling, specifically in link prediction.

While looking at the mobility data in aggregate, it is tempting to use the random walk model and its derivatives to approximate the mobility of people [BHG06]. Knowing, for example, the experimental probabilities of the destinations of an airport's outbound flights, one could say that any given individual at the location will choose one of the available destinations with its corresponding probability. However, at the individual level, these probabilities are very different, and cannot serve the purpose of modeling the next step of each person. Probabilities of destinations based on a history of a person's trips are much more indicative of that person's next step, as people are very likely to return to previously frequented locations [GHB08]. Therefore, it is necessary to model individual movement patterns, for example as described by Ashbrook et al. [AS03]. While Ashbrook used GPS as a location signal, the idea of individual Markovian mo-

bility prediction is platform-independent; hence, the findings can be applied in fixed-location sensing as well.

Another interesting approach is to exploit humans' predictability with respect to their social function, as the location of a person is often correlated with the location of their peers. De Domenico et al. [DDLM12] investigated the Nokia Mobile Data Challenge dataset[4] and achieved much better accuracy in future location predictions, if additional information about location of friends of a user were available, compared to the case where the prediction was only based on historical data of that user. While De Domenico does not claim generalizability of these findings, this does seem like a promising area of further research.

Link prediction does not deal directly with mobility, but does use mobility-derived features to approximate the development of dynamic networks, or recover missing data in static ones. Scellato et al. [SNM11] and Crandall et al. [CBC+10a] used location traces from online social networks to show that frequent co-location indicates higher probability of friendship links. Eagle et al. [EPL09] analyzed Bluetooth-based proximity traces and GPS locations of the observation participants to recover real-world friendship links. They found that co-location outside of working hours is highly indicative of mutual friendship, while work relations might result in asymmetric friendships. Crawnshaw et al. [CTH+10] presented a set of features characterizing co-location, like intensity, diversity, regularity, and specificity, and employed them to significantly improve the performance of link prediction. Linear regression, support vector machines and random forests are popular machine learning approaches applied in the link prediction problem.

In the following chapters I do not concentrate on mobility per se, but I do draw from the concept of socially bound movement patterns: I investigate groups of festival goers who share musical taste, and thus attend similar events in Chapter 2; in Chapter 4 I build on Cranshaw's work and propose more features describing different modes of co-location.

1.4 Privacy protection in computational social science

In the beginning of the data-collecting boom, the pioneering institutions were not fully aware of the potential of the information they gathered. Some of them released their datasets to encourage cooperation with the research communi-

[4] http://research.nokia.com/page/12000

ties, but these acts of goodwill led to a number of major privacy breaches. For example, in 2006 AOL Research released a list of web searches, along with numerical IDs of people who had performed them. It was not possible to recover the usernames, but the searches themselves contained personally identifiable information, which, cross-referenced with a phonebook, let The New York Times reporters discover the full identity of one of the users [BZ]. Other well known privacy breaches include the Netflix Prize dataset (de-anonymization of a user, along with her political, social and religious views described in [NS08]), and Group Insurance Commission medical records (recovering the medical history of then-Governor of Massachusetts from seemingly anonymous data described in [Swe01]. These examples show that the identity of participants can be recovered from very scarce and seemingly anonymized *breadcrumbs*. With the current trend in computational social science and the quantify-self movement of collecting much more sensitive data, and of significantly higher quality, another breach could lead to general distrust towards the new science and discourage people from participating in experiments.

Probably, the biggest challenge in protecting the privacy of observation participants is that it is impossible to predict what will be possible to do with the data in the future. A dataset, which at the time of releasing is prepared in a way that prevents de-anonymization with current methods, might be easy to break when more datasets are released and data from these multiple sources is cross-referenced. In order to prevent such events, the datasets are no longer published in a raw form with only the personally *directly* - identifiable information removed. The first step taken by many researchers is introducing the k-anonymity [Swe02], thus making sure, that in a response to a query against the dataset, any user is not distinguishable from k other users. However, as pointed out by Machanavajjhala et al. [MKGV07] this measure is vulnerable to *homogeneity attack*, where a particular feature is shared by k users and thus identifying one can leak information about others, and *background knowledge attack*, where additional information from other sources can be used to identify an individual in the set of k users. Therefore, Machanavajjhala proposed the ℓ-diversity measure, which ensures that the sets of users are diverse enough to prevent both types of attacks. After noticing that ℓ-diversity is susceptible to attacks targeting sets with attribute distribution significantly different than in the rest of the dataset, Li et al. [LL07] proposed the t-closeness method which addresses this shortcoming. k-anonymity, ℓ-diversity and t-closeness are based on binning users into sets of variable sizes and thus present a trade-off between the utility of the data and privacy of the participants: the smaller the sets, the higher the resolution of the data can be, and thus the researchers can draw stronger conclusions, but at the same time it is easier to identify individuals; on the other hand, the bigger the sets, the worse the resolution and less information for the researchers, but more privacy for individuals. So far, there is no agreement in the scientific community regarding adoption of a unified approach

to anonymization.

The datasets, which I have been working on during preparing this thesis, were not subject to any of the aggregate anonymization modifications. The directly identifiable information about each participant, such as their username, student ID number, email, and MAC addresses of their devices were scrambled, which prevented me from learning any specific details about the individual participants, potentially my peers, but would not prevent an ill-willed scientist from doing so.

1.5 Future of fixed-location sensing

Recent developments in mobile technologies make it increasingly convenient to move from fixed-location to mobile sensing. Indeed, this approach does not require infrastructure and will work wherever the devices are carried, including locations where setting up fixed sensors would not be possible. The main limitations of mobile sensing are that it requires will and consent of participants, and involves giving them relatively expensive hardware, thus narrowing down the number of possible users.

Fixed-location sensors on the other hand are significantly cheaper to deploy and can be used to observe people oblivious to the fact of the observation. While Bluetooth scanning has an unclear future due to the declining number of discoverable devices, there is no reason to expect WiFi adoption or mobile network development to slow down. Growing demand for mobile data transfers will push the providers to further extend the infrastructure, making the cell-tower based location more accurate. Still, the transfer speeds will not exceed those from WiFi connections anytime soon, so this technology will not be rendered obsolete, even for mobile users.

From yet another perspective, RFID might become *the* source of data about mobility of individuals: a single chip can be used as a credit card, public transportation ticket, library card, event access token, highway toll identifier, and many more. Since such chips can be embedded in mobile devices, I believe a smartphone will be the mobile sensing platform of the future, while still relying on fixed location infrastructure for many applications.

Chapter 2

Understanding Music Festival Participants' Behavior

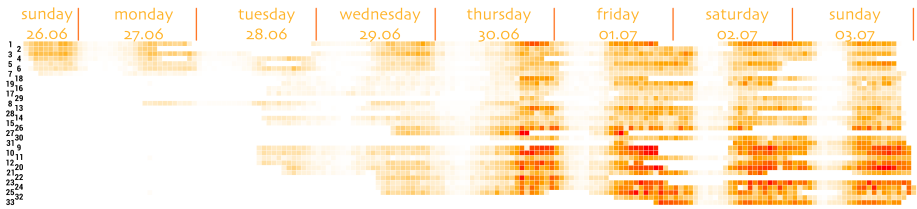


Figure 2.1: Unique Bluetooth devices observed throughout the 8 day festival by 33 proximity-based scanners, with the color intensity corresponding to the number of observations in one hour time windows. The scanners are grouped by stages and scanners at the main stages were deployed on day 4.

In this paper we present a study of sensing and analyzing an offline social network of participants at a large-scale music festival (8 days and 130,000+ participants). Spatio-temporal traces of participant mobility and interactions were collected from 33 Bluetooth scanners placed in strategic locations at the festival area to discover Bluetooth-enabled mobile phones carried by the participants.

We analyze the data on two levels. On the micro level, we use a community detection algorithm to reveal a variety of groups formed by the participants. On the macro level, we employ an Infinite Relational Model (IRM) in order to recover the structure of the network related to participants' music preferences. The obtained structure in the form of clusters of concerts and participants is then interpreted using meta-information about music genres, band origins, stages, and dates of the performances. We show that the concert clusters can be described by one or more of the meta-features, effectively revealing music preferences of participants. Finally, we discuss the possibility of employing the described method and techniques for creating user-oriented applications and extending the sensing capabilities during large-scale events by introducing user involvement.

2.1 Introduction

Mobile phones have become increasingly ubiquitous and an integrated part of our everyday life over recent years. This has led to a number of new possibilities in studies of human mobility, behavior, and interactions, as mobile phones can now be used to track people's activity. This area has recently received increased attention with studies of mobility by means of large phone data sets [SQBB10, GHB08] or sensor data collected on modern smartphones [EP06a, JLJ+10]. These studies have reported insights into fundamental human mobility patterns with results indicating very high levels of predictability.

In this paper we present a study of more than one hundred thousand music festival participants mobility, group formation, and music preferences at a large music festival in Denmark by using Bluetooth probing to discover mobile phones carried by festival participants around the festival area.

The use of Bluetooth technology as a way to gain insights into human behavior and mobility has also received increased attention recently [VDNVdW10]. Bluetooth technology has been applied in several different domains, and different schemes have been used. In a study of mobility by Hui et al. [HCS+05], participants were provided with a small active Bluetooth device that they were carrying throughout a conference to map participant mobility and events attended. Most commonly Bluetooth scanners have been situated in fixed locations to probe the presence of discoverable Bluetooth devices in proximity, which is also the approach presented in this paper. This method has been used for different applications including estimating the queue length expressed in waiting time in airport security areas [BHWS10, HAA+09]. Large scale studies of mobility by means of Bluetooth probing have also included tracking of vehi-

cles for the purpose of studying traffic patterns [HWB10], and large scale race events is another example [SLH+11]. In a related study O'Neill, Kostakos et al. [OKK+06, KOP+10] concentrate on the mobility and interactions of participants with regard to semantic meaning of locations where the Bluetooth scanners were deployed. They show profoundly different patterns of presence in places of different social function, for example a busy street vs. a bar. Unfortunately, even though they deployed more than 90 scanners, they only refer to four categories of locations - a street, the university entrance, an office and a bar. It is not clear how much insight they gained into the social structure of other locations.

In the context of human mobility in festival settings, a study by Versichele et al. [VNDVdW12] also applies proximity-based Bluetooth tracking to study mobility patterns. In their study 22 scanners were used over duration of 10 days with 1.5 million participants. However the general trend in their study is that participants only visit the festival short-term (typically one day), whereas the participants in this study are present at the festival area for up to 8 days, and select among 160 music concerts and multiple other events for the duration of the festival.

Where existing studies applying Bluetooth probing have focused on describing mobility patterns, this study involves a richer semantic context with information about concerts, music, genres, scenes, events, and participants, allowing a more detailed contextual analysis of participant behavior, and mobility.

More recently mobile sensor frameworks have been made available [API+11, KBD+10] enabling the collection of richer data sets capturing human behavior, mobility, and data for mapping social interaction through multiple channels. An advantage of having a mobile sensor framework on the smartphone is the potential in combining multiple sensor data to obtain finer granularity information and more robust estimations. For instance, data from sensors such as GPS, WiFi, GSM, and accelerometer can be combined to build a location estimator which works in different contexts (outdoors and inside the buildings) with higher accuracy than any single of these sensors can provide [MGP10]. However, a challenge in these studies is the deployment, which involves a mobile application running on participant devices. Therefore these studies have typically been carried out on a smaller selected population, but often over longer periods of time. As a result, the observational conditions and especially population sampling may introduce unknown biases. Although a mobile client (smartphone) may lead to very rich data sets, this methodology has a different set of challenges in terms of deployment at the festival. This includes supporting multiple clients and that participants have to actively install an application containing mobile sensing components. In this study the duration of the event is only 8 days, but using the Bluetooth probing technique we have access to a larger population.

In the following sections we describe the methodology, limitations, and challanges of data collection using Bluetooth scanning system in an environment with a limited and short-lived technical infrastructure. Next we present the data acquired during the 8 days of the festival and discuss the results of the Bluetooth discovery process. The chapter is concluded with a discussion of potential applications and the insights that can be obtained from studying the spatio-temporal data that can be acquired through Bluetooth probing.

2.2 Roskilde Festival

Roskilde Festival is one of the six biggest annual music festivals in Europe and is held south of Roskilde in Denmark. It started in 1971 and since 2009 has been attracting more than 100,000 participants annually (with up to 30% being volunteers). In 2011 it gathered an estimated 130,000+ people. The festival lasts for 8 full days, starting Saturday evening and finishing Sunday at midnight. For the first 4 days only the camping grounds and a small festival zone are open, including a single stage (Pavilion Junior) featuring up-and-coming Nordic bands. On Thursday afternoon the main grounds are opened, the major music events start and last for the next 4 days.

The main festival grounds cover about 0.2 km^2 with 6 stages of various sizes. The festival campsite, located south of the main festival grounds covers nearly 1 km^2. In addition to the stages, the grounds include cultural zones, shops, restaurants, artistic installations etc. Participants can freely move through the grounds in the daytime; once the concerts are finished for the day, the main grounds are closed and then open in the morning hours next day. Some areas in the main grounds are off limits for participants, such as backstage areas or technical areas behind merchandise passages.

In 2011 the participants consisted of 77,500 festival guests, around 3,000 press representatives, 3,000 artists, 30,000 volunteers, 20,000 one-day guests plus an unknown number of guests over 60 and under 10 years old – we estimate that at least 130,000 people were present at the festival during the 8 days in total. 54% of the population were women and approximately 22% of the audience visited the festival for the first time. The average age was 23 years and a typical participant was a student living in a Scandinavian city. About 80% of the participants came from Denmark, 8% from Norway, 4% from Sweden, and 8% from other countries[1].

The six stages host concerts of different sizes and genres:

[1]Source: http://roskilde-festival.dk/

- Orange stage, capacity 60,000+, all genres
- Arena stage, capacity 17,000, all genres
- Cosmopol stage, capacity 6,000, hip-hop, electronica, urban world music
- Odeon stage, capacity 5,000, mixed, mostly rock
- Pavilion stage, capacity 2,000, mixed, mostly rock
- Gloria stage, capacity 1,000, mixed, experimental

2.3 Methodology

Our study of human mobility in the festival settings relies on discovering Bluetooth-enabled devices that are operating in discoverable mode. As Bluetooth is a short-range low-power protocol for implementing Wireless Personal Area Networks (WPAN), it limits the range in which Bluetooth-enabled devices can be discovered. It operates on the Industrial, Scientific and Medical (ISM) frequency band of 2.4GHz [PBK06]. Communication always happens in master-slave mode and is established between new devices with a master device sending inquiry packets to discover nearby devices that are in the inquiry scan substate (discoverable). Discoverability of a device commonly needs to be set manually by the user, and can be either limited in time or set to infinite. It is worth noticing that for instance Android-based smartphones (until recent versions) only allow time-limited discoverability, while iOS devices (iPhone, iPod, etc.) and WindowsPhone smartphones are only discoverable while the user is interacting with the Bluetooth menu. While this limits the number of potential phones we can discover significantly, we show that there are still many discoverable devices.

In the present study Bluetooth scanners functioned as master devices, broadcasting inquiry messages (scanning) continuously. Responses from the devices in proximity were silently logged, without any active participation on the user side. This is similar to the approach described in [HH09] where tracking of the individual in a non-invasive way is considered more suitable for large-scale studies. The received signal strength intensity (RSSI) of the response was not registered. Although it is technically possible to use RSSI to calculate the position of the discovered device through multilateration [Ben07, Kel10], the accuracy of the approach varies depending on the environment. Moreover, due to the limited range of Bluetooth, we considered position accuracy obtained from a single scanner (i.e. around 10 meter radius for class 2 Bluetooth devices) sufficient.

2.3.1 Bluetooth Scanner Device

Off-the-shelf Nokia N900 smartphones were used as Bluetooth scanners with custom software built for detecting Bluetooth-enabled devices in proximity. Off-the-shelf hardware was used as a relatively simple solution, providing 3G communication (necessary for obtaining the results in real time from the large festival area), data storage, battery power (for the events of short power outages), GPS for tracking the device in case it was lost, and finally a Bluetooth module. The data from the scans was stored in a local SQLite database on the device and additionally uploaded to a server, depending on the network availability. Scanner and uploader applications were running on the smartphone, and extra background processes restarted them if required. This was to ensure the highest possible availability and robustness of the system.

A scan for discoverable devices typically takes about 30 seconds, so scanning performed as frequently as possible results in approximately two scans per minute. Devices that did not upload data to the server for a prolonged period of time were rebooted either by issuing a command via Bluetooth or by manually turning them on and off. In order to minimize this effect, periodical reboot every 24 hours was enforced in the software.

The collected data is a time-series of events. Each of the events is described by the time, scanner ID and a Bluetooth MAC address of a discovered device. This information does not enable us to link the device to the person (such as name or personal identification). Thus, the Danish Data Inspectorate considered the information handled in this project as being non-sensitive information about the participants thereby enabling the observations to be made without special permissions or requiring informed consent from the participants. To ensure that not even the detected devices were identifiable afterwards, the MAC addresses were hashed after extracting information about the vendor. The human-readable identifiers (Bluetooth friendly names) of the devices were not retrieved in order to improve the scanning time and to ensure anonymity of the participants.

2.4 Observational Study

The data was captured through 33 Bluetooth scanners placed in strategic positions around the festival site, as shown in Figure 2.2. The scanners were placed in the vicinity of the stages, as those were the most interesting, semantically rich spots. However, since the availability of power sources was crucial while choosing the exact location, and the infrastructure at the festival is only tem-

porary, the scanners were mainly located in the shops, beer booths (close to the counters) and mixing areas of the stages. Those locations provided sufficient coverage of relevant areas to discover patterns in participants' mobility.

Figure 2.2: Map of Roskilde Festival inner area with indication of the location of Bluetooth Scanners. The orange areas indicate places for the audience for the respective stages.

The Bluetooth scanning data was uploaded in real time via the 3G network for real time processing purposes, but upload of the data was of course subject to network availability. Problems with the mobile network connections were occurring due to a high number of mobile phones in a relatively small physical area, especially during large concerts. Therefore, for some scanners the collected data was uploaded once the connection was available (typically in early morning hours). 7 of the 33 devices were running without manual intervention for the whole period of the festival, but the rest had to be maintained one or more times during the festival. For instance, if the power had been switched off for more than about 7 hours (typically in the early morning) the devices had to be manually turned on.

The radius of Bluetooth is limited to about 10 meters for the transmitters used in most of the mobile phones (class 2). While this makes it possible to pinpoint the location of the observed devices, it also makes it a challenge to collect representative data in a large area, as it will only be partly covered. The devices observed by a scanner could belong to a person only passing by; on the other hand, a person staying right outside the radius of the coverage even for the whole concert might not be discovered.

Figure 2.3: Nokia N900 Bluetooth scanner in a protective box attached under a beer booth counter.

2.5 Data Collection and Analysis

The deployed Bluetooth scanners collected 1,204,725 observations during the 8 days of festival activities. This included a total of 8,534 unique devices discovered, meaning an average of 141 observations of each device during the festival. Overall, this corresponds to at least 6.5% of the population at the festival having been observed in the study, thereby providing a window to understand festival participant behavior, mobility, and interactions.

Table 2.1 provides an overview of the observations from the 33 scanners used in the study. As can be seen from the table the most unique devices were discovered by scanner 9, 10, 23, 24, and 25 that are all located around the largest stage where most participants would be expected to be seen. Each of those scanners discovered more than 4,000 unique devices. An overview of unique devices observed throughout the 8 days of the festival is shown in Figure 2.1 on the first page.

Beyond serving as a unique identification of the device the MAC address is structured so the vendor of the device can be determined from first three octets of the address (24 bits) formally known as an "Organizationally Unique Identifier" (OUI) [IEE]. The list of the assigned OUIs is managed by IEEE, designated by the ISO Council to act as the registration authority. Some identifiers found in the devices may not correspond directly to the end-product manufacturers, as they may be registered under subcontractors company. In total, around 70

#	Obs	Uniq	O	#	Obs	Uniq	O
1	77145	3607	1	18	28844	2302	3
2	44224	1880	9	19	32773	2245	8
3	53706	3091	11	20	34264	4753	15
4	31836	1801	15	21	22022	3473	20
5	33167	3265	16	22	20003	1901	2
6	38834	2120	20	23	43784	4372	29
7	28440	1102	3	24	53695	4404	27
8	40648	893	0	25	55025	4429	51
9	49852	4316	2	26	61706	3290	12
10	45813	4116	28	27	24714	1900	16
11	21714	3467	3	28	32512	1651	8
12	30027	3433	31	29	27944	1491	5
13	60276	2770	11	30	32067	2411	22
14	34202	3159	8	31	15616	2514	21
15	36293	2582	5	32	19190	2553	22
16	22044	1809	3	33	25578	2934	18
17	20280	1227	13				

Table 2.1: An overview of the 33 scanners with numbers of observations (Obs), and unique devices per scanner (Uniq), and unique devices only discovered per scanner (O). Total number of unique devices was 8,534.

unique vendors were discovered, however the 7 largest vendors account for 96% of all unique devices and 99% of all observations, as shown in Figure 2.4.

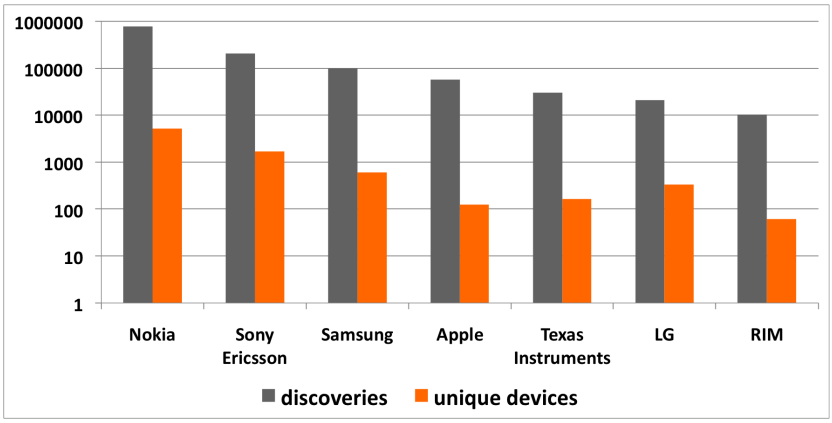


Figure 2.4: Number of unique observations and devices (log scale). The 7 largest vendors account for 96% of the devices and 99% of the observations.

2.6 Modeling

One of the interesting questions regarding events such as music festivals is the internal structure of the crowd: whether people move alone or in groups and how groups are different. In addition, the influence of music taste on collective group decisions on concert selections is interesting as is the mobility of the groups. In attempt to understand what insights on these issues we can gain from the data obtained in the presented study, we analyze the data at two levels. Firstly, we concentrate on the micro level by running a community discovery algorithm. Then, we investigate the macro level to determine the general trends of attendance and relate the findings to the available meta information regarding the schedule of the festival and types of artists.

2.6.1 Micro Groups Modeling

We understand micro groups as sets of people frequently co-occurring in spatio-temporal bins. We divide the timeline of the entire festival into 1076 x 10-minute temporal bins and 10-meter radius areas around the scanners create the spatial

bins, as shown in Figure 2.2. A similar technique of inferring social links from spatio-temporal co-occurrences is described in [CBC+10b].

Out of 8,534 discovered devices we rejected those that were seen in fewer than 10 temporal bins or fewer than 3 spatial-bins. Those devices were considered belonging to participants for whom we do not have sufficient data or being stationary devices (such as crew laptops). After this processing 5,339 devices were obtained (63%) and it is for these devices that we calculated. For all these common co-occurrences were calculated. The weights of the links were calculated as the number of co-occurrences of participant A with participant B divided by total number of occurrences of participant A (A to B edge). This creates a directed graph, where A can be important to B but not necessarily the other way around. This accounts for the asymmetry in the participants' activity and different natures of their relations. For the visualization and subsequent analysis, only links that occurred in at least 3 different locations and weight at least 0.5 (seen in 50% of all observations of the participant) were chosen.

The final constraints on the discovered micro groups are strong: they require that from 130,000 participants, people are seen within 10 minutes in a radius of approx. 10 meters at least half of the times they are observed in total and in at least 3 different locations, to ensure sufficient entropy for meaningful modeling. The constraint are imposed on existence of each edge, hence the directed edges. This should ensure that the discovered motifs are in fact people moving around together. It was found that 12 nodes (devices) were forming structures (pairs and square) with perfect correlation of occurrences, which we consider devices belonging to the same person. Based on the discovered groups, a directed graph can be constructed, with edges indicating the discovered friendships. In total, 574 nodes with 448 edges were detected. The motifs can be seen in Figure 2.5. The most interesting are the structures with high connectivity, indicating groups of participants observed to often move together.

The baseline for the micro group detection was calculated using rewiring algorithm [MSZ04], shuffling the participants in spatio-temporal bins. For N=35 tests $\mu = 5$ nodes and $\mu = 3$ edges were discovered ($\sigma = 4.32$ and $\sigma = 2.60$ respectively). This indicates that the recovered structures are not an effect of random movement of participants but reflect an actual underlying structure.

The star structure visible in the upper left corner of Figure 2.5 with multiple inbound edges and none outbound is an interesting artifact showing a person working in a shop in an area covered with several scanners. The person was frequently picked up by 3 scanners (1, 2, and 3) with customers also picked up there but independently from each other. Similar artifacts were seen in larger number when the threshold of common co-occurrences was set to 2, since some of the long beer booths had two scanners placed in them. Such star structures

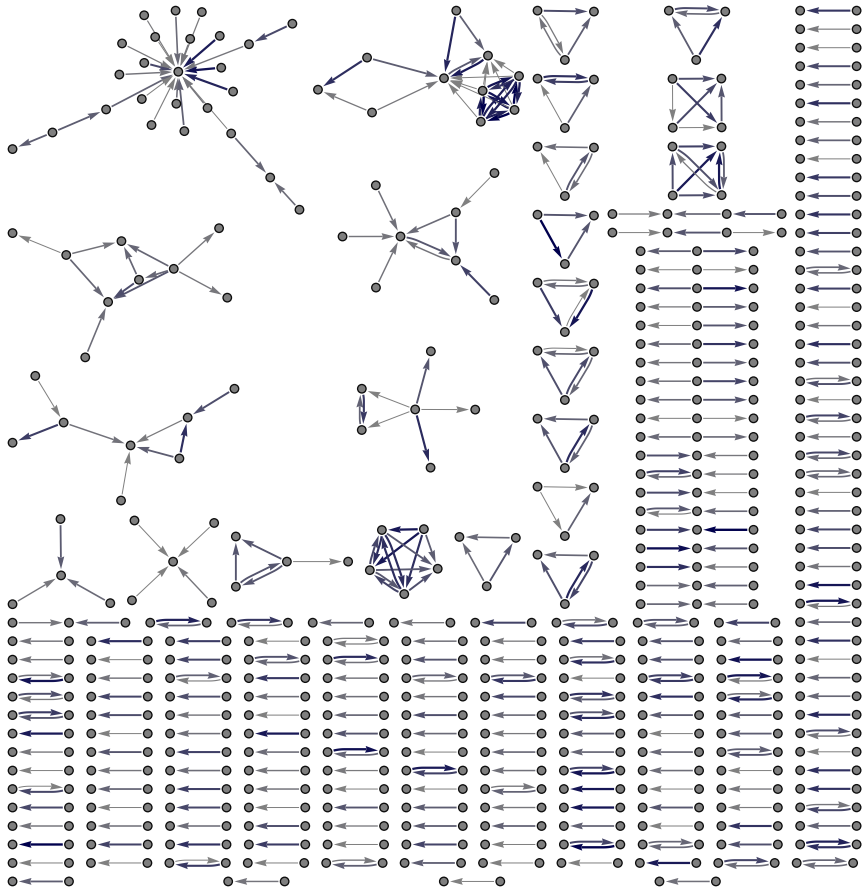


Figure 2.5: Directed graph of discovered micro-groups of participants. Participants frequently seen together create couples, triangles and larger structures, providing insights into the internal structures of the festival crowd.

with inbound links can be summarized by saying that those places (represented by people working in them) were important to participants, but participants were not significant for them.

The presented algorithm for detecting micro groups and discovered structures is a simple example of possibly very granular analysis of the collected data. With extremely small spatio-temporal bins we still recover over 500 people moving around while belonging to a particular structure.

Size of stage	Small	Medium	Big
ρ	0.2462	0.0351	0.3427
$P-Value$	0.0333	0.8593	0.0091

Table 2.2: Correlation between popularity of the band (log playcount) and the number of unique devices

2.7 Macro Groups Modeling

We combine the spatio-temporal traces with the band schedule, to find out which concerts each of participants attended. Next, we assign a set of meta information to each show. This way we establish a richer semantic context and analyze the guests' motivations for choosing particular concerts. The metadata consists of:

- *genre* – based on available Last.fm tags, each band is manually assigned with one genre label from the following: *electronic*, *rock/pop*, *folk/world*, *hip-hop/rap*, *metal/punk/hardcore*, *other*
- *playcount* – number of times Last.fm users listened to music of a band
- *country of origin* – from the Roskilde Festival schedule; the countries have been grouped into following categories: *Denmark*, *Other Nordic*, *USA*, *Western Europe*, *Other*
- *scene* – from the Roskilde Festival schedule
- *date* – from the Roskilde Festival schedule

Intuitively, the number of people at the concert would be highly correlated with the intensity with which people listen to the bands, i.e. the playcount. To verify this assumption, we calculate Pearson's correlation between the number of unique devices found during each concert and the logarithm of playcount of the band, see Table 2.2. We group the concerts according to the size of the stage they performed at. As shown in Table 2.2 there is a small (if any) positive correlation between the popularity of the band and the number of discovered devices. This shows that people's choices regarding the concerts they attend cannot be fully accounted for in this way and more complex modeling should be used to reveal more interesting patterns.

2.7.1 Data pre-processing

Our Bluetooth traces are a time-series of events, each of which contains the participant id, scanner id, and time. The goal of the pre-processing stage is to transform the behavioral time-series data into a binary attendance table, which maps each participant to the concerts she attended. In each event, we assign the scanner to the stage where it was located. Then, we assume that scans which took place between 10 minutes before the starting time of a concert and 1 hour 45 minutes after that moment were taken "during" this concert. Thus, we determine during which concert, if any, each event happened. This results in a matrix where each element represents the number of times each participant was scanned at a given concert. To indicate whether a given participant actually attended a concert, we transform the table to a binary table by setting a threshold on the number of observations.

2.7.2 Outlier detection

The binary table created in pre-processing contains two categories of outliers. Firstly, there are guests who participated in less than three concerts and are thus irrelevant in terms of the analysis. Bluetooth devices, which were recorded throughout the festival at the same location such as employee cell phones or laptops at a particular stage constitute the second category of outliers. These are defined as entities which participated in at least 70% of concerts at one stage and at least in twice as many concerts at that one stage compared to all the other stages in total. After removing outliers, 5127 attendees are left for further analysis.

2.7.3 Metadata pre-processing

We obtain the community assigned tags for each band from Last.fm. There are more than 400 unique tags associated with the participating bands and for our modeling purposes we need to significantly reduce the dimensionality of this data. Based on the most significant tags and manual verification, we assign each band to one particular genre: *electronic*, *rock/pop*, *folk/world*, *hip-hop/rap*, *punk/metal/hardcore*, *other*. Such categorization is, of course, highly simplified, but provides a satisfactory representation of kinds of music performed at the Roskilde Festival.

2.7.4 The Infinite Relational Model

We fit an Infinite Relational Model (IRM)[KTG+06, XTYK06] to the binary attendance matrix to reveal the underlying patterns of people's behavior at the festival. Note, that the Model is oblivious to the accompanying meta information such as genre, band's country of origin, date, and location of each show.

The IRM is a model for binary relational data (graphs) and can be characterized by the following generative process for bipartite graphs. First, each of the row and column nodes is assigned to a cluster according to the Chinese restaurant process (CRP). The CRP is an analogy for building a partition ground up by assigning the first node (i.e. customer in a restaurant) to a table and subsequent nodes (customers arriving at the restaurant) to an existing table, i.e. cluster, with probability proportional to how many existing customers are placed at the table and at a new table, i.e. cluster, with a probability proportional to the parameter α. Customers thereby tend to sit at most popular tables making the popular tables even more popular – an effect commonly described as, "the rich get richer". The partition of the nodes induced by the CRP is exchangeable in that the order in which the customers arrive does not influence the probability of the partition[Pit06]. Next, link probabilities are generated which specify the probability of observing a link between clusters; and finally, the links in the network are generated according to these probabilities. For bipartite graph we have the following generative process:

$$\boldsymbol{z}^{(1)} \sim \mathrm{CRP}(\alpha^{(1)}), \qquad \textit{Row cluster assignment,} \tag{2.1}$$

$$\boldsymbol{z}^{(2)} \sim \mathrm{CRP}(\alpha^{(2)}), \qquad \textit{Col. cluster assignment,} \tag{2.2}$$

$$\eta_{\ell m} \sim \mathrm{Beta}(\beta, \beta), \qquad \textit{B/t. cluster link prob.,} \tag{2.3}$$

$$A_{ij} \sim \mathrm{Bernoulli}(\eta_{z_i^{(1)} z_j^{(2)}}), \qquad \textit{Link.} \tag{2.4}$$

Inference in the IRM model, i.e. determining the posterior distribution of the cluster assignments, entails marginalizing over the link probabilities, which can be done analytically. This is a major advantage of the IRM model, enabling inference by Markov Chain Monte Carlo (MCMC) sampling over the cluster assignments alone. Marginalizing over link probabilities, i.e. $\boldsymbol{\eta}$, we obtain the following joint posterior likelihood

$$
\begin{aligned}
p(\boldsymbol{A}, \boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)} | \beta, \alpha^{(1)}, \alpha^{(2)}) = & \quad p(\boldsymbol{A} | \boldsymbol{z^{(1)}}, \boldsymbol{z^{(2)}}, \beta) p(\boldsymbol{z}^{(1)} | \alpha^{(1)}) p(\boldsymbol{z}^{(2)} | \alpha^{(2)}) \\
= & \quad \left[\prod_{\ell m} \frac{\mathrm{Beta}(N^{+}_{\ell m} + \beta, N^{-}_{\ell m} + \beta)}{\mathrm{Beta}(\beta, \beta)} \right] \times \\
& \quad \left[\frac{{\alpha^{(1)}}^{L} \Gamma(\alpha^{(1)})}{\Gamma(I + \alpha^{(1)})} \prod_{\ell=1}^{L^{(1)}} \Gamma(M^{(1)}_{\ell}) \right] \cdot \\
& \quad \left[\frac{{\alpha^{(2)}}^{L} \Gamma(\alpha^{(2)})}{\Gamma(J + \alpha^{(2)})} \prod_{\ell=1}^{L^{(2)}} \Gamma(M^{(1)}_{\ell}) \right],
\end{aligned}
$$

where $L^{(k)}$ is the number of clusters, $M^{(k)}_{\ell}$ is the number of nodes in the ℓth cluster of mode k, and $N^{+}_{\ell m}$ and $N^{-}_{\ell m}$ are the number of links and non-links between nodes in cluster ℓ and m. Using Bayes theorem the conditional distribution of the cluster assignment of a single node is given by

$$
\begin{aligned}
p(z^{(1)}_i = \ell | \boldsymbol{A}, \boldsymbol{z}^{(1)} \backslash z^{(1)}_i, \alpha^{(1)}, \boldsymbol{z}^{(2)}, \beta) &\propto \left[\prod_{m} \frac{\mathrm{Beta}(N^{+}_{\ell m} + \beta, N^{-}_{\ell m} + \beta)}{\mathrm{Beta}(N^{+\backslash i}_{\ell m} \beta, N^{-\backslash i}_{\ell m} \beta)} \right] q^{(1)} \\
p(z^{(2)}_j = m | \boldsymbol{A}, \boldsymbol{z}^{(2)} \backslash z^{(2)}_j, \alpha^{(2)}, \boldsymbol{z}^{(1)}, \beta) &\propto \left[\prod_{\ell} \frac{\mathrm{Beta}(N^{+}_{\ell m} + \beta, N^{-}_{\ell m} + \beta)}{\mathrm{Beta}(N^{+\backslash j}_{\ell m} \beta, N^{-\backslash j}_{\ell m} \beta)} \right] q^{(2)},
\end{aligned}
$$

such that $q^{(k)} = \begin{cases} w^{(k)}_{\ell} & if\ w^{(k)}_{\ell} > 0 \\ \alpha^{(k)} & otherwise \end{cases}$ where w_{ℓ} is the number of nodes already assigned to cluster ℓ and $N^{+\backslash i}_{\ell m}$ and $N^{-\backslash i}_{\ell m}$ denotes the number of links and non-links between nodes in cluster ℓ and cluster m not counting any links from node i of mode one (j is similarly used to denote not counting any links from node j in mode two). Hence, a new cluster is generated according to the CRP with probability proportional to $\alpha^{(k)}$. By (Gibbs) sampling each node assignment of the row ($z^{(1)}_i$) and column ($z^{(2)}_j$) clusters in turn from the above posterior distribution we can infer $\boldsymbol{z}^{(1)}$ and $\boldsymbol{z}^{(2)}$. The inference thereby also estimates from data the number of groups in each mode.

We note that this posterior likelihood can be efficiently calculated only considering the parts of the computation of $N^{+}_{\ell m}$ and $N^{-}_{\ell m}$ as well as evaluation of the Beta function that are affected by the considered assignment change. Notice, the expected value of the relations $\boldsymbol{\eta}$ given the node assignments $\boldsymbol{z^{(1)}}$ and $\boldsymbol{z^{(2)}}$ is defined by $\langle \boldsymbol{\eta}_{lm} \rangle = \frac{N^{+}_{lm} + \beta}{N^{+}_{lm} N^{-}_{lm} + 2\beta}$. Apart from the above Gibbs sampling we also include so-called split-merge moves to improve the inference [JN04]. The split merge procedure was implemented with three restricted Gibbs sampling sweeps initialized by the sequential allocation procedure of [Dah05]. *Infinite*

relational model can be efficiently applied to large datasets using GPU computing [HMH11], which could allow for real time applications. Here we set $\beta = 1$, $\alpha^{(1)} = log(I)$ and $\alpha^{(2)} = log(J)$, where I is the number of unique devices and J is the number of concerts.

2.7.5 Robustness of the model

We use a number of measures to evaluate the generalizability of the results and robustness of the model. The model estimation procedure is run 110 times; each time 2.5% of the links and an equal number of non-links are treated as missing, and then used for prediction. Firstly, normalized mutual information (NMI) is calculated between each pair of estimated models. Notice, $0 \leq NMI \leq 1$ where 0 indicates no relationship between the two assignment matrices and 1 indicates a perfect correspondence [HMH11]. The NMI scores for the concert assignment matrices average at 0.91 with the standard deviation of 0.03, while the score for the attendee assignment matrices has the mean of 0.45 and standard deviation of 0.02. The relatively low NMI for the clusters of participants is related to the fact that the model forces the assignment of each attendee to only one cluster. There can be many such assignments which are equally valid and thus with every run of the model calculation the final participant groups vary. Since the assignments of concert clusters are significantly more stable, they will be in focus of further analysis.

The predictive performance of the model is measured using the Area Under Curve (AUC) of the Receiver Operator Characteristic. AUC evaluates how well the distributions of links and non-links are separated. Notice, $0 \leq$AUC≤ 1 where 0.5 indicates separation not better than a random guess and 1 indicates a perfect separation. This measure is not vulnerable to class imbalance problem [TSK06]. The average value of AUC for the 110 models is 0.81 with the standard deviation of 0.01. Finally, it is shown that after 150 iterations the log probability of the model converges to a stable value across 110 runs, see Figure 2.6. It is important to emphasize that this stability is achieved for the models trained on non-complete datasets (with each run 2.5% of links and the equal number of non-links were randomly discarded to be used for prediction). As shown in Figure 2.6 the model is robust to random initialization conditions as well as to data partially missing.

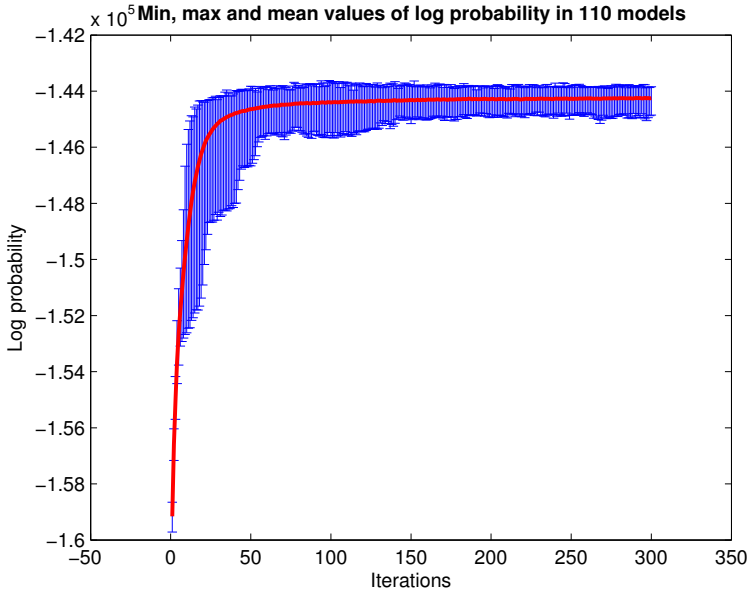


Figure 2.6: Robustness: Independently of random initialization conditions and parts of the data used for cross-validation, the final value of log likelihood is stable across 110 trained models.

2.7.6 Results

After having proven the stability and generalizability of the used method, more models are calculated based on the full attendance table, without treating any part of the data as missing. The model with highest log probability is used for further investigation. As shown in Figure 2.7 this model groups 5127 people in 16 clusters and the 160 concerts in 25 clusters. The color-coded value of η indicates the between-cluster link probability. In subsequent sections these values are interpreted and related to the available meta information.

2.7.6.1 Relating chosen concert clusters to available metadata

This section describes particular findings, which further justify the use of the chosen technology as well as provide additional insight into the audience dynamics.

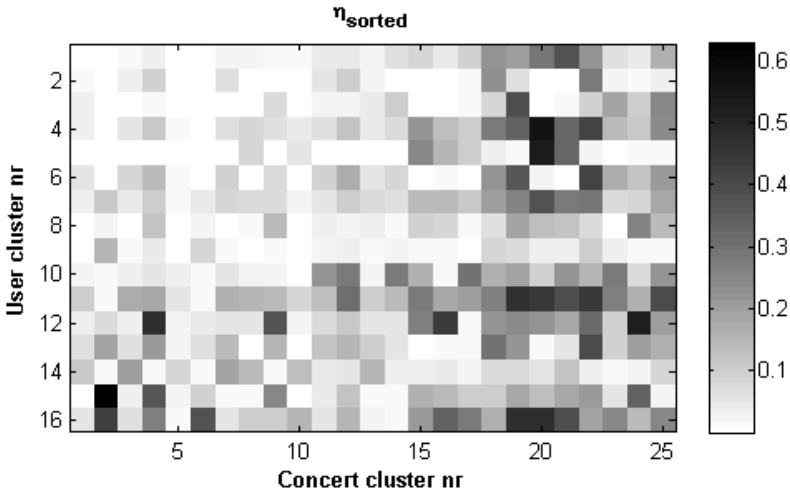


Figure 2.7: Between cluster link probability for the estimated 16 clusters of attendees and 25 clusters of concerts, with clusters sorted by size in descending order. Preference regarding the choice of concerts can be observed, for example user cluster 5 is strongly associated with concert clusters 15, 16, 17, 20, 21 – many people in cluster 5 attended concerts from these clusters.

Figures 2.8 - 2.12 show the distribution of concerts in the created clusters in relation to particular features. We only consider first 10 clusters, containing between 24 (cluster 1) and 7 (cluster 10) concerts. This captures .725 of all concerts at the festival. With fewer concerts in clusters, it is increasingly hard to provide meaningful interpretation.

We use χ^2 test to compare the distributions in the clusters against the overall distribution to understand if the cluster bears any meaning in relation to the particular feature. It should be however emphasized that the results are not rock-solid: with such a small number of concerts in the clusters, the results are more of guidance in relating the clusters to available metadata, rather than quantification of the findings. Still, we can note that the model produces interpretable results, giving insight into the festival structure.

Figures 2.8 - 2.11 show the distribution of concerts from clusters (1-10) across the available meta information. The last column in each figure indicates whether the distribution in that cluster is significantly different than the overall distribution: if yes, the cluster can be considered meaningful and explained by this feature.

Figure 2.8 shows that the clusters are quite structured in terms of the dates. It is intuitively understood - festival participants present at that particular day attend these concerts. As shown in Figure 2.9 only two clusters have distribution of genres different than overall distribution. These two clusters clearly point to electronic and folk/world genres. Figure 2.10 deals with the distribution of origin of the bands and shows three clusters with well-pronounced grouping of the bands: Danish, Danish+Nordic, and USA.

Figure 2.11 indicates that most the clusters display strong grouping of the bands based on the stage where they happened. This may be related to the fact that concerts of similar type (if not necessary the same genre) are planned at the same stage; also, participants' mobility is limited and a common behavior of participants may be to stay at the same stage.

The summary shown in Figure 2.12 makes it clear that the model produces clusters primarily based on the stages where they took place. Interestingly however, we also see the influence from the date of the concert, origin of the band, and the genre. Although the presented results are not very strong statistically, we conclude that the model does produce clusters that relate to features of the concerts/bands.

We can describe the produced clusters (1-10) based on their relations to features:

1. Electronic concerts from the main days of the festival, happening at the three stages (Cosmopol, Gloria, Odeon).
2. Danish bands playing in the warm-up days at Pavilion Junior stage.
3. Various genres from the first days of the main festival from three stages (Cosmopol, Gloria, Odeon).
4. Concerts from the first days of the main festival from Pavilion stage.
5. Mainly concerts from the second (largest) day of the main festival from various stages.
6. Danish and other Nordic bands entirely from the warm-up days.
7. Folk and World bands from the main days of the festival, mostly from the smallest Gloria stage.
8. Bands from the US playing various genres on the last day at different stages.
9. Various bands from the main days playing at Pavilion.

10. Concerts happening on the last day, possibly capturing one-day-ticket participants.

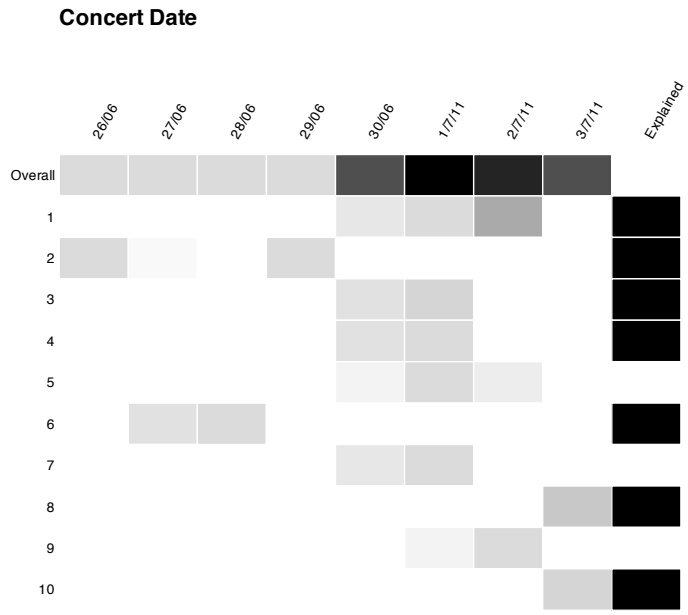


Figure 2.8: Distribution of concert dates in clusters. 7 out of 10 clusters have dates distribution significantly different from the overall.

2.7.6.2 Between cluster link probability matrix

As shown in Figure 2.7, there are several clusters of participants, which show very specific preferences regarding the concerts. For example, participant group 5 (392 persons) only attended concerts from clusters 8, 10, 15, 16, 17, 20, 21. Nearly all of the concerts in these clusters took place on 3rd of July (last day of the festival with major bands performing). Participants from group 4 (475 persons) showed similar preference on that day but they also attended concerts on other days. Participant group 6 (352 persons) behaved like participant group 4 on days other than 3rd of July but showed no interest in the concerts on that day. Another participant group, which shows a clear pattern in concert

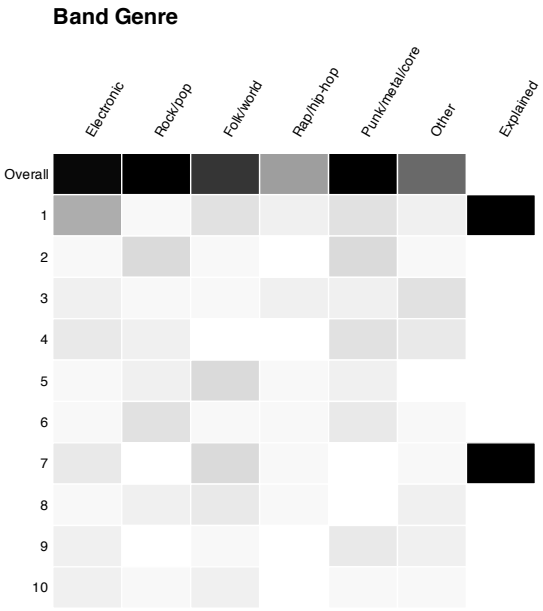


Figure 2.9: Distribution of concert genres in clusters. Two clusters have distribution significantly different from the overall.

attendance, is group 12 (91 persons) which has high link probabilities with clusters 4, 9, 16, and 24. It occurs that all of the concerts from these clusters took place at the Pavilion stage.

2.8 Discussion

Our study has demonstrated that discovery of Bluetooth devices at large-scale events can provide interesting insights on participant behavior, group formation, and music preferences. The analysis of the collected Bluetooth data has demonstrated how the spatio-temporal data can reveal underlying structures, when combined with additional contextual metadata describing the concerts and music genres. In the present study we found that over the duration of the festival 6-7% of the participants appear to have Bluetooth switched on and in

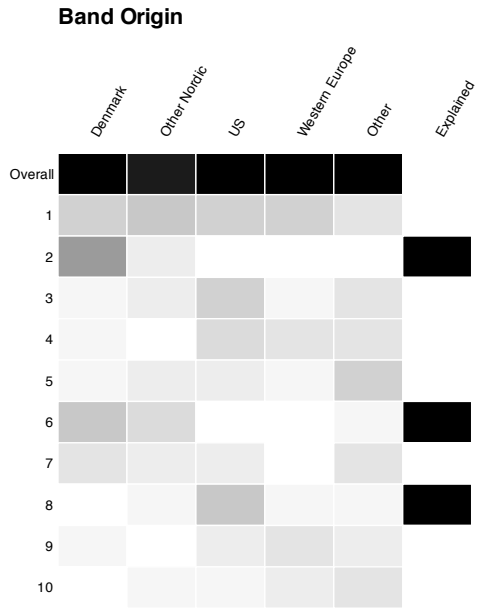


Figure 2.10: Band origin distribution in the clusters. Three clusters show significant grouping of bands: Denmark, Denmark + Other Nordic, and US.

discoverable mode. However, based on the available data it is not possible to conclude on the reasons for this or the actual usage of Bluetooth. Moreover, we were able to observe the distribution of vendors of the discovered devices, but this distribution may not correspond directly to the actual distribution of mobile phones at the festival. In other words, the Bluetooth discoverable devices may not be representative, as for instance most Android-based smartphones only allow time-limited discoverability. As such we would expect to observe fewer Android devices in our dataset than there actually are at the Festival. The increasing adaptation of the Android smartphones could perhaps account for the lower penetration of Bluetooth-discoverable devices in the crowd when compared to [VNDVdW12].

The spatio-temporal data allow for analysis of co-occurrences of participants, thereby giving indications of group formation among the festival participants.

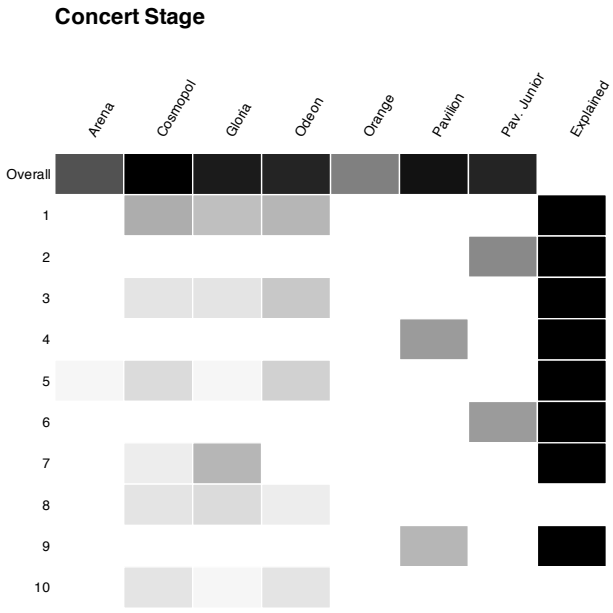


Figure 2.11: Distribution of concert stages. We can notice most of the clusters displaying significant grouping of the concerts according to the stage.

Furthermore, an advantage of the Bluetooth methodology for doing participant census is that we learn the identity of devices. With this, it is not only possible to estimate the number of people present at different concerts but also determine patterns in the selection of different concert across the entire festival, based on music profiles determined from the spatio-temporal data. Therefore the analysis of this data have provided insights into the underlying structures, that is, the discovery of groups with specific behaviors (music preferences) in terms of choosing concerts. Our analysis shows, that the allocation of artists in terms of stage and day of Festival when they perform is a crucial issue. We find that many people are not willing to move around the festival area - instead participants tend to spend much of their time around a particular stage. We also show, that for those who do attend concerts at different locations, the country of origin of a band are important factors when selecting the gigs. Furthermore, we do not find clusters of fans of particular music genre ,which means the participants are open

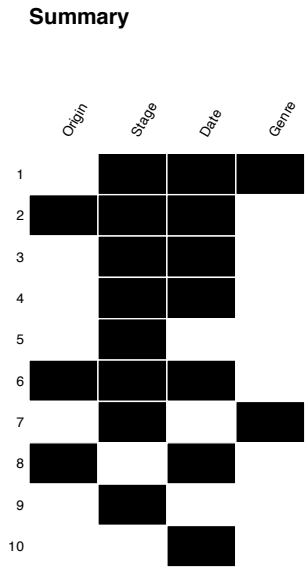


Figure 2.12: Summary of the clusters and features where their distribution is significantly different from overall.

towards different kinds of performances. Such information can be very valuable for the Festival organizers in the process of booking and allocating bands to stages.

As the collected data was uploaded continuously by the scanners it was possible to create a near real-time visualization of the location of participants at the festival. The real time visualization displayed the activity as the number of unique devices seen in half-hour time windows in different zones of the festival and mapped this information onto a 3D model of the festival area. The rotating 3D model was displayed on a 46 inch monitor located in the so-called *Social Zone* of the festival and ran in continuous loops, displaying speed up of activities from the first day until the current moment, see Figure 2.13. This way of visualizing the activity data allowed for high dynamic of normally slower changing patterns, an easy overview of the festival activity so far, and the possibility of incorporating past data that was only uploaded later (in case scanners did not

have a network connection). This setup also allowed us to test the feasibility of obtaining the Bluetooth data in real time using the regular cellular 3G network as a way to build end-user applications on top of the system.

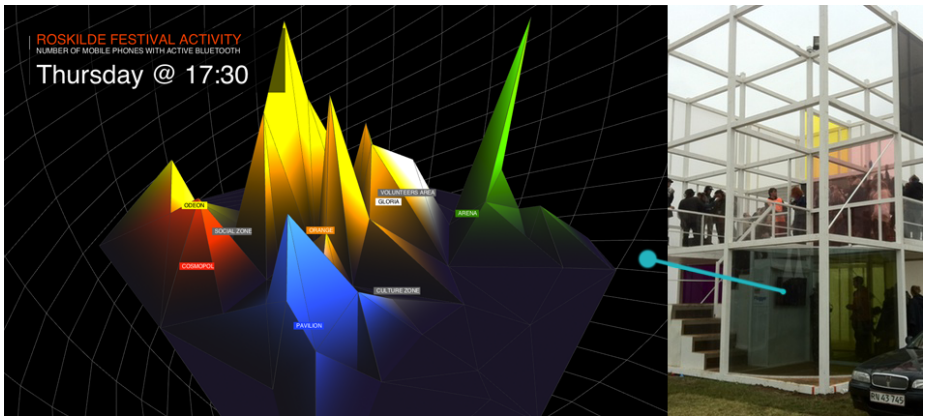


Figure 2.13: The 3D real time animated visualization shown to participants on a large display situated inside a cubic installation that also hosted a silent disco. The 3D model of the festival area was continuously rotating and replaying the visualization of the collected Bluetooth data from the beginning of the festival up to the current moment.

At the festival we were able to observe participants as they experienced the visualization of the Bluetooth data. Initially, they were attracted by the animation, bright colors, and high dynamics and then they subsequently understood what was shown in the visualization. In the setup that was deployed at this festival, the interaction through the 3D visualization of the Bluetooth devices in the festival areas was indirect. The analysis of the data has demonstrated that even more sophisticated participant feedback could be included in such a visualization – even in real-time. Furthermore, it could allow for more direct interaction through mobile social apps on participant smartphones. For instance to locate groups, participants, or relevant events, as they are happening at the festival.

As mentioned in the introduction, sensor frameworks for smartphones have received increased attention recently. Future studies could further improve the data collection at a large-scale event through the richer datasets that can be obtained from smartphone embedded sensors. By distributing the scanning on multiple client devices the inherent limitation of the present short-range proximity based probing approach may be addressed. In the current setup it is challenging to cover a large physical area in addition to the set of challenges in

deploying the system – including limited availability of power and network in the festival settings. However, a challenge in the distributed scanning approach is the deployment of a sufficient number of client devices in order to obtain sufficient continuous coverage of the area. The initial steps in the direction of distributed Bluetooth scanning were taken by Stopczyński et al. [SLL+13].

We believe that the results that can be obtained from this Bluetooth probing methodology may also be useful on multiple levels for the festival organizers. The data can help the organizers in assessing participant reactions to the music selection and distribution over the different stages. A more detailed analysis of participant mobility may also help the organizers in planning the layout of the festival area for future festivals.

2.9 Conclusions

We have shown that proximity-based Bluetooth sensing is a useful method for obtaining spatio-temporal data in a large-scale event setting. It is possible to analyze the data, accounting for sparsity and missing data using mathematical models and discover meaningful patterns of participant behavior, including mobility, group formation, and music preferences. We have also demonstrated the feasibility of capturing Bluetooth data from a large crowd and visualize the resulting spatio-temporal data in real time. Finally, we have proposed how the Bluetooth probing methodology may serve as a framework for creating future mobile social interaction applications for such large-scale events.

2.10 Acknowledgment

We would like to thank the Roskilde Festival organizers. Also thanks to Nokia for partly sponsoring the mobile phones used as part of the study. Finally thanks to Krzysztof Siejkowski, Marcin Ignac, and Søren Rosenbak.

Chapter 3

Detecting Face-to-Face Meetings Through Fixed-Location Sensors

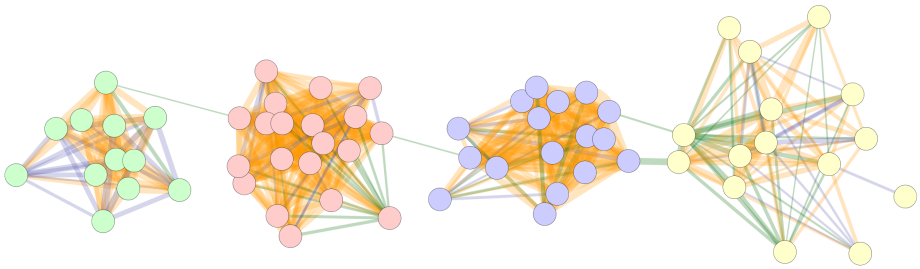


Figure 3.1: The network composed of 466 most active dyads, where color nodes denote different study lines. 366 of them are among most active both on and outside of the campus (orange edges), while 50 are only significantly active on (blue edges) and 50 outside of the campus (green edges). The campus-centered subset is a good approximation of the whole network, but note that the bridging edges between distinct communities occur only outside of the campus

How do we detect social interactions in computational social science? Using curated Bluetooth traces collected during a large computational social science study deployment (N=130 participants) as a ground truth for face-to-face meetings. This study reports on a thorough examination of the data in search of quality deficits, biases, and interplay between the information channels. We argue that this to be crucial step, before the data can be used to generate and verify scientific hypotheses. As our primary focus, we examine how well the Bluetooth signal can be recovered from other channels, such as cell tower traces and a campus-wide WiFi system. Our main result is that interactions, which occur on campus (and can thus be discovered using the system WiFi), constitute a relevant approximation of all the interactions among the participants. This finding may be beneficial for 1) discovering co-location networks in contexts, where a WiFi system is already deployed, and using additional mobile devices is not feasible, such as university and company campuses, schools, and other institutions, 2) planning more energy-efficient deployments of experiments involving mobile devices, and 3) understanding how different signals may introduce bias into analysis and conclusions.

3.1 Introduction

This paper is an initial overview of the longitudinal data about 130 students which we gathered in the first months of the SensibleDTU[1] project deployment. The data was collected through mobile devices carried by the participants using a customized version of Funf framework [API+11], as well as via campus-wide wireless network system. Thorough examination of the data in search of quality-deficits, biases, and interplay between the information channels is an important first step to be taken before the data can be used for generating and verifying scientific hypotheses. We use detection of face-to-face meetings as an example application and verify how various channels can be used to approach the issue of validation.

Face-to-face meetings are used in computational social science as a prominent signal for discovering social ties. Different methods have been used for recording such meetings, for example video analysis [Ros02], infrared sensors [OWK+09], or more recently Bluetooth scanning using mobile phones [EPL09], but little work has focused on thoroughly quantifying and comparing the utility and precision of different methods.

[1] http://www.sensible.dtu.dk

3.2 Related Work

To our best knowledge, none of the large scale computational social studies to-date has produced a comprehensive evaluation of the utility of various methods of F2F (Face-to-Face) meeting detection. As pointed out by Cattuto et al. [CBB+10], studies, which span millions of users lack in temporal and spatial granularity, while deployments with high accuracy are difficult to scale up. They propose sensing the F2F contacts by purpose-built RFID (Radio-Frequency Identification) badges, which can send and receive packets using several discrete power levels, to control the distance of detectability. Additionally, since human body absorbs the frequency the bages use, they sense only in the direction the user is facing. Olguín et al. [OWK+09] presented a similar approach, as they employed purpose-built Sociometric Badges with IR (infrared) transceivers and microphones, which are able to detect the presence of other Badges within 1-meter distance inside a 30° cone. Choudhury [Cho04] showed how the autocorrelation of sound captured by two Badges could be used to augment the signal from the IR sensors, resulting in a method, which achieves an 87.5% recall of conversations lasting longer than one minute (experimental results are evaluated against manually assigned labels).

While using custom built RFID or IR badges can provide researchers with precise estimations of social networks emerging from F2F contacts, these methods are only applicable in specific contexts. Since the badges must be worn on the outer layer of clothing, preferably at a fixed height, one may only expect acceptance from specific audiences who are required to wear some kind of badges anyway, for example in companies or during conferences. Capturing interactions across many different contexts is more efficiently facilitated using less obtrusive devices. Additionally, using off-the-shelf hardware can significantly decrease the deployment cost and enable researchers to scale up the projects.

Increasingly ubiquitous smartphones provide users with additional utility compared to purpose-built Badges, and thus increase the probability, that the users will remember to wear the sensors on a daily basis. The Near Field Communication (NFC) chips available on today's smartphones require a maximum distance of only a few centimeters [NFC13], and are therefore not applicable for detecting F2F contacts. Other sensors are typically employed, for example Bluetooth. Liu and Strigel evaluated the utility of Bluetooth as a proxy for face-to-face interaction [LS11] and point out that with a range of 10 meters, BT visibility might not necessarily imply face-to-face contact, as these usually occur at significantly smaller distances. They propose an RSSI (Received signal strength indication)-based model augmented with light sensor data for classifying proximity in different contexts. In a controlled environment, they classify proximity events as those occurring within 1.5 meters with accuracy higher than 80%.

Using WiFi receivers is even more difficult due to sensor's lower spatial accuracy with a maximum range of up to 250 meters in outdoor settings. Kjærgaard and Nurmi [KN12] describe challenges related to using on-the-device WiFi for social sensing, pointing out noise sources as well as poor generalizability of machine learning inferred features for detecting face-to-face meetings. In another study, Kjærgaard et al. [KWRT12], focused on detecting indoor pedestrian flocks by centrally analyzing data collected by mobile WiFi sensors. They show high recall and accuracy in sensing groups and their findings are evaluated by ground truth based on video recordings.

Carreras et al [CMSO12] proposed a different approach, which exploits the ability of Android smartphones to function as Portable Hot Spots. In their experiments, phones run software, which switches the WiFi mode between receiving and transmitting thus allowing devices to achieve bi-directional discovery. While their claim a spatial resolution of 0.5 meters, this accuracy is based on modeling RSSI; a strategy, which is susceptible to differences in environment. Moreover, the method requires devices to be in an 'awake' state at all times, which prevents them from using normal WiFi transmissions. Therefore this strategy is not applicable in real-life deployments.

Finally, Banerjee et al. [BAB+10] introduced a system, which uses multiple radios (BT and Peer-to-Peer (P2P) WiFi), as well as an empirically determined model for RSSI-to-distance function (to minimize the distance estimation error), arriving at $0.9m$ median accuracy. Their other contribution is to propose an adaptive sampling system, which leverages social aspect of scanning. The presented method is not directly portable to operating systems other than WM 6.0, but the notion of P2P communications for exchanging proximity data is a novel insight.

A majority of other studies within computational social science concentrate on a single source of information to detect F2F meetings, for example GSM towers [WPS+11, GHB08, EP06b, RZZB12], Bluetooth [EP06b], or WiFi AP visibility [KH04, JLB05] referring to neither the ground truth (actual face-to-face meetings which did occur) nor data from other sensors.

3.3 Methodology

3.3.1 Bluetooth as a proxy for physical proximity

Firstly, we quantify the quality of Bluetooth as a proxy for physical proximity. In order to achieve this purpose we run a small experiment with three phones scanning the Bluetooth environment every five minutes. In our setup, we ensure that distances between the phones are not well below five meters, with all three phones in the same room, so they should all detect each other with each scan. However, we observe that the experimental probability of a phone being detected by another device in this setup varies from 0.828 to 0.895, depending on the device, but irrespective of the distance.

We introduce 10-minute time bins for the remainder of the experiment. We assume a face-to-face meeting between two persons, if there were at least two positive discoveries between the phones they carry, regardless of the direction. Using the worst-case scenario detection probability of 0.828, we find that the probability of the system failing to detect dyadic meeting is lower than 2% (assuming independence between the discovery events).

Among 130 participants we chose 65 for whom we have the results of at least 80% of all possible Bluetooth scans during a period of three months. We divide the experiment timeline into 9504 10-minute bins and create a 3D array with dimensions user $\times$ user $\times$ timebin, which indicates whether two users met in the time bin (1), did not meet (0) or we lack the data to determine whether the meeting happened (-1). There are 175044 dyadic meetings in this matrix. In the remaining analysis we focus on the bins and dyads for which we have data and define this setup as 'ground-truth' information to compare other channels to.

3.3.2 System WiFi

System WiFi is an interesting source of information. If we can show that system WiFi can be used to reliably infer social interactions, this implies that any campus, company, or urban wireless network can be used as a social sensor with no need for additional hardware, or active involvement of the participants. On most smartphones Bluetooth discoverability is disabled, rendering them unvailable for Bluetooth scanning. All Android [And13], iOS [App13], BlackBerry [Bla13] and WindowsPhone [Win13] devices ship with discoverability turned off. Combined smartphone market share of these devices was estimated to 96.9% in

4Q12 by IDC [IDC13], while the smartphone market penetration in Denmark was estimated to 45% in 1Q12 by Google [Goo12]. The same devices, however, will automatically connect to known WiFi networks without any action from the user. As a consequence WiFi routers, used as sensors may provide highly useful social data (given proper privacy precautions, such as informed consent, etc.). Such information can be used to determine the social context of an event, for example how many people are present at a location at a given time, with much higher accuracy than Bluetooth alone. It is therefore crucial to examine the quality of the data system WiFi produces.

In our system there are 500+ access points (APs) providing WiFi around the campus area and their location is known. Students typically use their unique credentials to access the network and so it is possible, from the system perspective, to determine the AP to which each student is connected, and, as a consequence, in which building they are located. The system reports lists of students connected to each access point every 5 minutes. We integrate this information over time to create time bins spanning 10 minutes and corresponding to the Bluetooth scans.

We assume co-location between users in two separate cases: 1) whenever participants are connected to the same AP, and 2) whenever they are connected to APs mapped to the same building.

3.3.3 GSM towers

The cell probe is configured to collect the cell tower information every 10 minutes. The cell ID are not mapped to a physical location, so users connected to the same physical tower will appear as connected to different towers if they use different providers. While it surely is a shortcoming, most mobile network dataset in use only deal with data (an thus users) from a single telecom provider with limited market share, and are thus even more limited. In our dataset information in 12% of person-timebins is missing. We define a co-location as two users connected to the tower with the same cell id and area code in a time bin. After comparing co-locations to the Bluetooth-based ground truth, it occurs that only 10% of all meetings are correctly detected and 97% of all positives are false. Due to very low precision and recall of this method, we do not consider cell id as a way to detect face-to-face contact in more detail in this paper.

3.4 Analysis

3.4.1 Campus as a proxy for the world

In this section, we attempt to answer the question, whether the social activity on the campus is a good approximation of the social activity between the students in general. Firstly, we verify, that being connected to the system WiFi is an acceptable proxy of physically being on campus, using location data collected through GPS receivers and network-based location estimators running on the participants' phones. For each time bin when at least one of their location estimations falls within the campus bounds, we check whether they are connected to the system WiFi. As shown in Figure 3.2, there are only two persons, who do not use system WiFi at all, while all others use it extensively, with a median probability of 71%.

Next, we analyze whether the social interactions, which can be observed through system WiFi, are representative sample of all the interaction, which happen between the participants. Figure 3.4 shows that, in general, the time dyads spend together on campus is highly correlated with the time they spend outside of campus ($\rho_{on \leftrightarrow out} = 0.78$, $p_{val} = 0$) and even higher with the total interaction time ($\rho_{on \leftrightarrow total} = 0.93$, $p_{val} = 0$). This holds more for dyads, which interact much ($\geq$10 hours during the experiment - $\rho_{on \leftrightarrow out} = 0.51$ and $\rho_{on \leftrightarrow total} = 0.81$), than for dyads with less interaction ($\rho_{on \leftrightarrow out} = 0.28$ and $\rho_{on \leftrightarrow total} = 0.60$).

Now, we closely analyze the most active dyads to find out how well the network emerging from meeting outside of campus is represented by the network visible trough system WiFi. As shown in Figure 3.3a, the dyads, which are most active on campus, are not the same as the dyads most active outside. However, as we approach top ~22% dyads in both categories, overlap approaches the ideal case, while the Kendall's τ rank correlation coefficient is equal to $\tau = 0.63$, $p_{val} = 0$. The ~22% of top dyads consists mostly of links between people from the same study line, with very small number of significant intra-study line dyads. As we look into the overlap of more than top ~22%, it approaches random, which indicates that interactions which appear to happen between students from different study lines might just be random proximity events, for example in cafeterias, with little social meaning. It also occurs, that the ~20% most active dyads account for ~90% of all interaction in the ground truth dataset, both on and outside of campus (see Figure 3.3b). This can be explained by the fact, that lectures, which the students take together, generate a lot of BT "interaction". With some students not being connected to WiFi, some interactions happening during classes are perceived as "outside of campus" and dominate the emerging network.

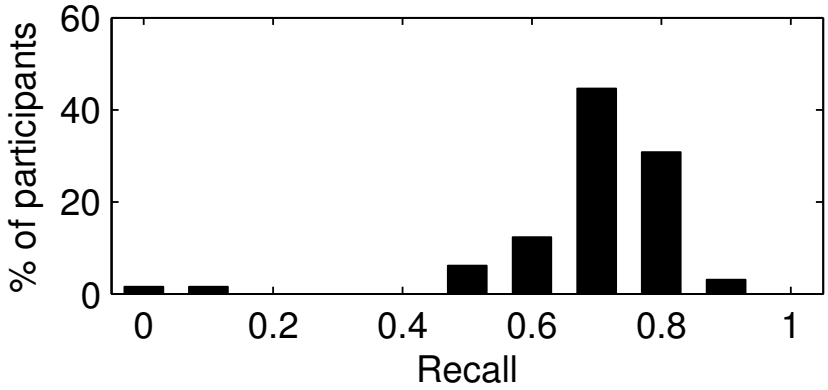


Figure 3.2: Most of the participants use system WiFi extensively, with just a few preferring 3G while on campus

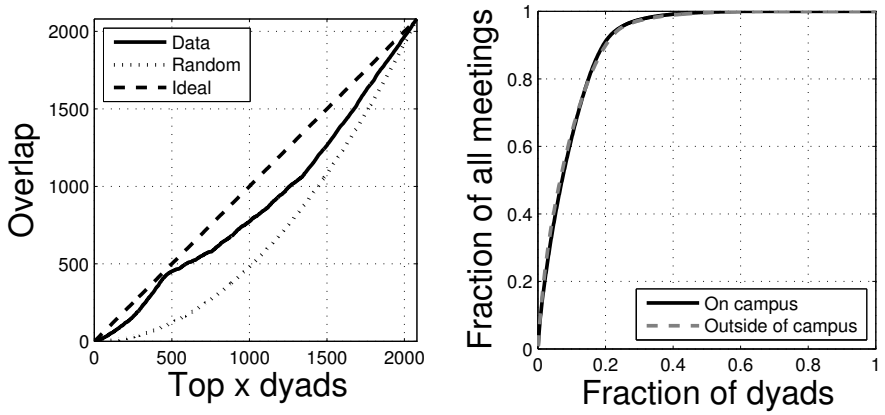
Figure 3.1 shows the visualization of networks consisting of 20% most active dyads in both categories.

3.4.2 System WiFi

As described in Methodology, we can chose to assume co-location whenever two persons are connected to the same AP ('same AP method'). Or when individuals are connected to APs mapped to the same building ('same building method'). By creating matrices corresponding to the Bluetooth ground truth matrix, we discover the following:

- 44% of all ground truth meetings occur while both parties are connected to system WiFi
- 94% of these meetings happen when both parties are detected in the same building, and 64% while both persons are connected to the same AP
- 47% of dyads in the same building, and 62% of dyads connected to the same AP meet face-to-face according to the BT channel

Clearly, using the same AP method provides higher precision but lower recall, while using same building method allows us to recall nearly all of the meetings, but also produces many more false positives. Inside most buildings on campus, there are multiple APs visible to the users, so even users who are in close



(a) **(b)**

Figure 3.3: (a) The dyads, which are most active on campus, are not the same as the dyads most active outside. However, as we approach top ~22% of dyads in both categories, the overlap approaches the ideal case. Then again, the distribution approaches random; (b) 20% of dyads in the dataset account for 90% of all interaction time and the distribution is virtually the same for meetings on and outside of campus

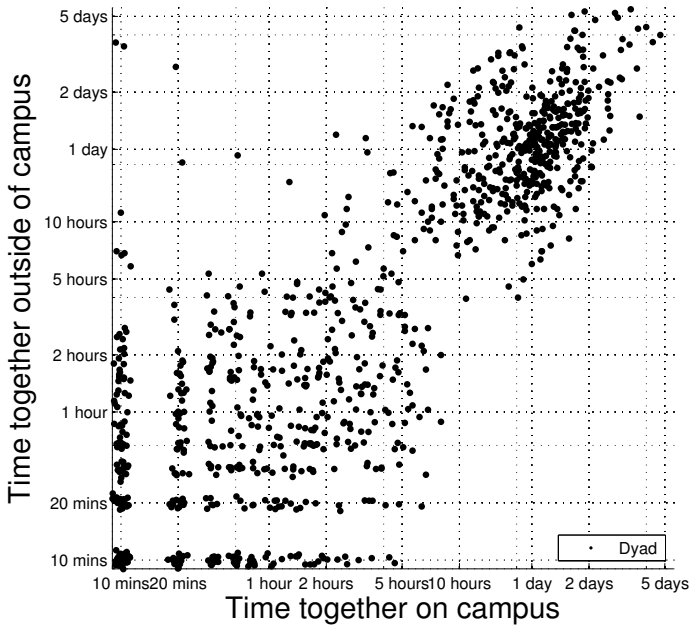


Figure 3.4: The time dyads spend together on campus is correlated with the time they spend outside of the campus. Two groups seem to emerge: those that spend below 10 hours together and those that spent more. Random noise has been added to the times below 5 hours to better visualize the number of dyads.

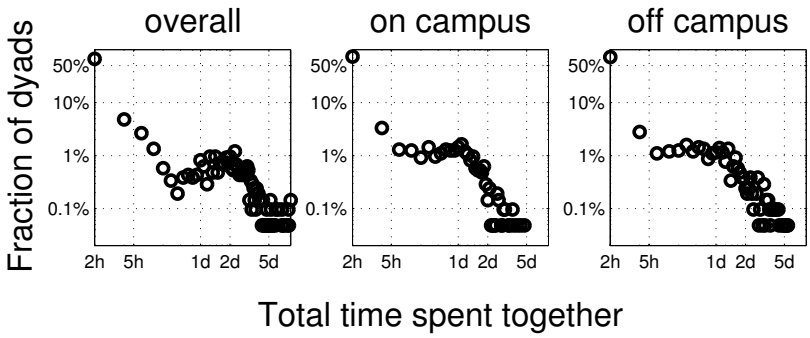


Figure 3.5: The figure presents the histograms (2 hour fixed duration bins) of time spent together by dyads. In the overall distribution, two groups can be easily observed: dyads that hardly spend any time together, and dyads that spend much time together.

proximity might be connected to different routers. On the other hand, some buildings are extensive and have several semantically different spots inside, for example a dining hall and lecture halls. To further analyze the results, we create the map of the campus with buildings color-coded according to their precision, see Figure 3.6. It is clearly the case that not all the locations around campus are equally reliable when it comes to producing accurate results.

Inspecting the map, we learn that buildings with more traffic produce many more false positives than buildings which are less populated. This indicates, that any algorithm for recovering face-to-face meetings from the system WiFi signal should take the total number of persons present into account. Furthermore, buildings, which are close to bus stops, provide very low precision (South-West, South-East corners and the center of the map).

3.5 Summary of results

We have used Bluetooth visibility as a ground-truth proxy for physical proximity. We show, that even though a single Bluetooth scan is likely to fail to detect some devices within proximity-range, we can minimize the risk of false negatives by introducing 10-minutes time bins, as well as by assuming reciprocity. Via this

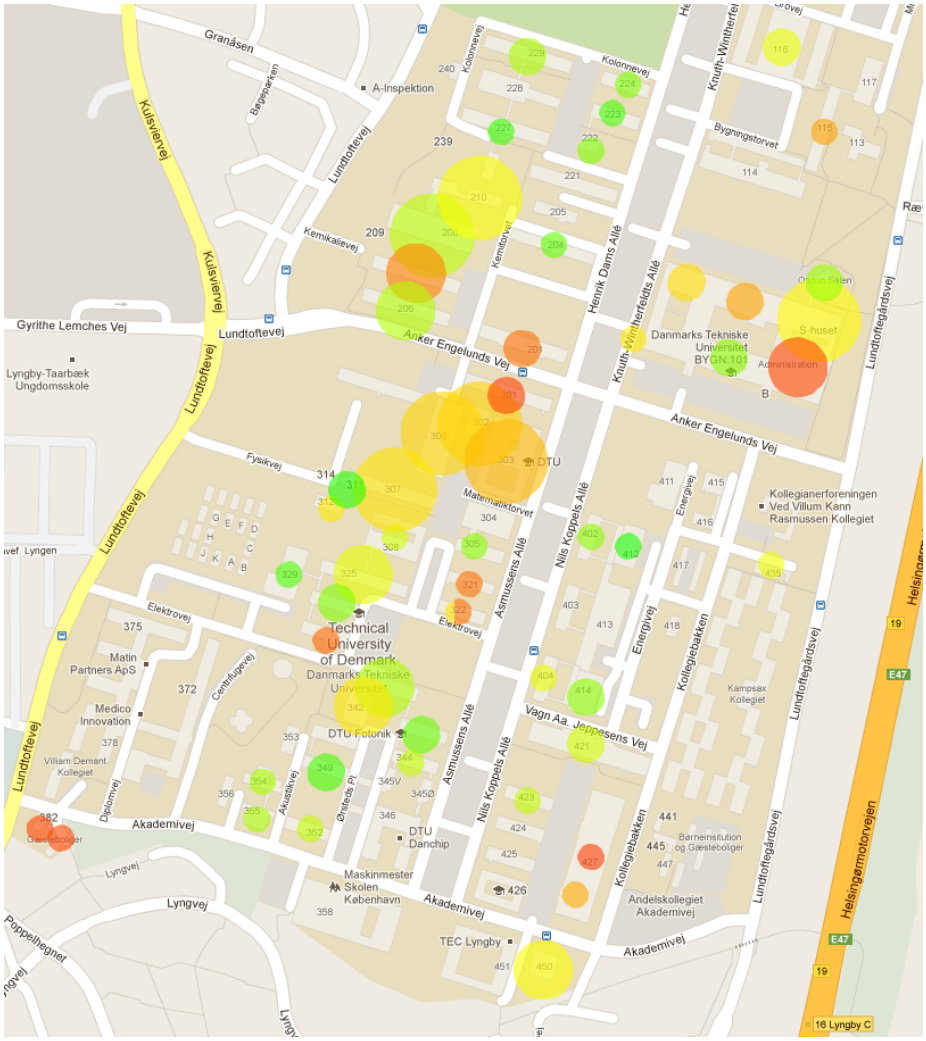


Figure 3.6: Campus buildings are color-coded according to the precision of their face-to-face meeting detection through system WiFi. Green color denotes perfect precision, yellow - 50% precision, and red - 0% precision. Size of the node is correlated with count of dyadic co-locations which occurred in each building during the experiment

relatively simple definition of a face-to-face contact we arrive at only ~2% of false negatives.

Furthermore, we showed that social interactions, which happen on campus, are

a good approximation of the social life between the participating students in general. Only very few dyads meet exclusively on or outside of campus, while most of the active users reveal similar habits in both contexts. Therefore, the social networks constructed using most active dyads in both contexts are highly overlapping.

Finally, we show that a campus-wide WiFi network can be used to recover a large part of social interactions, which occur on campus. Thus, with a few reservations, a WiFi system may act as a viable alternative to expensive deployments of sociometer experiments, which require active participation from users and purchases of sensing hardware.

3.6 Future Work

This abstract merely provides a glimpse into the comparison between Bluetooth-based ground truth and the campus-centered system WiFi perspective.

As shown in the section on previous work, Bluetooth itself is not a flawless proxy of F2F contacts, and the first step in future work will be a thorough analysis in search of more accurate ground truth. Next, a sensible balance to the precision/recall trade-off in recovering F2F meetings from system WiFi should be proposed. One possible solution would be to merge some APs into common semantic locations, on a sub-building level (e.g. based on machine learning). The APs for merging could be chosen among those, to which users are connected while being perceived as "together" through the Bluetooth channel.

Since system-perspective WiFi is by design likely to produce a significant number of false-positive interaction detections, machinery for extracting meaningful events will also be necessary. For example, two persons connected to the same AP might not be having a F2F meeting, but if they are spotted often and/or in different location, these events should be considered more important.

Furthermore, two very rich channels, location and user-perspective WiFi have not yet been fully explored here. These two channels compliment each other, each providing context in different scenarios. GPS-based location works best outdoors, where WiFi coverage might be poor. On the other hand, user perspective WiFi can not only provide location estimation where GPS signal does not reach, but it also has a potential of pin-pointing face-to-face meetings, and determining static vs. dynamic contexts if there is a sufficient number of APs available. Finally, once dependencies between the information channels are better understood, it will be possible to propose a model for recovering face-to-face

meetings, which are currently *'lost'* due to missing data and based on available input from other channels.

Chapter 4

Recovering real friendships from spatio-temporal traces

Until recently, social scientists have not had access to objective methods to observe behavioral patterns and interactions of large populations. It has been shown that people are not very good at recalling frequency of interactions with others accurately or even fail to sort their peers by amount of time they spend together [BKS79]. Thus, modeling the "objective" social networks based on interaction was impossible. Since this situation started changing, computational social scientists developed a number of models based on the actual social networks of physical proximity, for example to predict the spread of diseases [BCG+09]. However, it can be argued that for example information spread depends much more on the cognitive social networks than those emerging from counting physical contact: two people in proximity will exchange viruses spreading through droplet route even if they do not know each other, but they will not exchange news unless there is a acquaintanceship between them. This is why it is important to revisit the cognitive, self-reported networks of acquaintanceship and friendship. In this paper we analyze what behavioral patterns can be extracted from physical proximity events and how they can be used to reconstruct the social networks in the form, in which people perceive them.

4.1 Introduction

Sociologists have long argued that one of the key factors in bond formation is physical and/or psychological proximity between people, also referred to as propinquity[Kra]. Physical co-location was indispensable in the past, due to limited possibilities of contacting people in distant places. However, with the development of the Internet and online social networks such as Facebook it has become much easier to "meet" like-minded persons without meeting them in person. One could imagine that physical proximity has lost some if its importance. In this paper we attempt to evaluate how informative this factor is for link prediction in a context of a group of tech-savvy, Facebook-using students at a major European University.

Over the last few years we have witnessed significant growth in usage of mobile electronic devices such as smartphones, tablets, and laptop computers. To fully exploit the potential of these devices, one needs to be constantly connected to the Internet. Virtually all the students and faculty members used the Wi-Fi network at our campus extensively. Software, tracking the number of people connected to each wireless access point, was developed to estimate the usage of the network and improve the quality of service. Due to the fact that anonymized versions of usernames are stored as well, we can trace the mobility patterns of individuals and derive an offline social network from co-occurrences between persons.

4.2 Related work

Much work has been done to demonstrate how analysis of location traces can be applied in the problem of link prediction in social networks. Eagle et al.[EP06b, EPL09] used software running on mobile phones of 100 participants to track the users' location as well as proximity. Then, through eigenvalue decomposition, they found social routines and locations, which were indicative of different types of relationships. For example, acquaintances co-occur on weekdays during office hours more often than friends. However, off-campus meetings, especially during weekends and outside of office hours prove to be highly indicative of friendship. The behavioral data collected using the mobile phones was compared to self-report information acquired through surveys. Crawnshaw et al.[CTH+10] described their analysis of 500 users' check-ins on an online location sharing network. The check-ins were done automatically every 10 minutes without the necessity for the user to explicitly state their location. The coordinates were acquired from GPS and by looking up positions of available Wi-Fi routers. They

proposed a number of features characterizing both locations and users to facilitate the inference of friendship from co-occurrences. Scellato et al.[SNM11] examined the temporal evolution of another online location sharing network and concentrated on the friends of friends of a user. This way, they narrowed the prediction space significantly and made it possible to analyze the vast dataset (380 000 users). They suggested further location-based features to further improve the performance of the link prediction. By acquiring the data from the location sharing online networks, Scellato et al. were able to map the geographic coordinates to particular venues along with their accompanying meta-information (office building, restaurant, disco, etc.). They have shown that whether two people form a friendship link is highly correlated with them visiting the same locations. In another research, Crandall et al.[CBC+10a] used geo-tagged pictures uploaded by users to a photo-sharing website. Instead of venues, they introduced equal-area partitioning of the globe in a similar fashion as Crawnshaw. In their data, due to the extensive area in question (the whole Earth) even a very small number of co-occurrences are highly indicative of friendship.

The presented works highlight a trade-off between the resolution as well as quality of the collected data and the number of participants in an experiment. Eagle et al. collected their data without active participation of the users at around 5 minutes resolution, but this research required purchasing and servicing hardware for all subjects, thus limiting their number. Crawnshaw, in turn, built an application, which runs on both mobile phones and computers, but the extent of their research was limited by the number of people who opted in for sharing their location constantly. On the other hand, Scellato et al. were able to access the data about few thousand times more people, but with irregular time sampling, as the requirement for people to manually check-in at locations greatly reduced the number of samples.

In most of the mentioned works based on social networks, the friendship is defined as the link between two users of the social network. Thus, contrary to data acquired through self-report, as in[EPL09], all the relationships are symmetric. Moreover, while there is no constraint on how many friends one can have in an online social network, there is a biological restriction to how many peers people can have a meaningful relationship with, as pointed out by Dunbar[Dun92]. Therefore, profiles in online social networks can be more connected than real-life persons they represent.

In our study we analyze location traces of 7700 people with a fine resolution of 10 minutes. Additionally, we obtained self-reports from two disjoined groups of students. Due to the fact that their location data is aggregated by access points around the university campus, we have no information about the location of participants when they are outside of the campus or whenever they are not using the Wi-Fi network. However, as we show further in the paper, the information

we have collected is sufficient to effectively predict majority of real, self-reported friendships in the students' social network.

4.3 The Dataset

In this study we combine information from several sources as described in this section.

4.3.1 The Wi-Fi network data

4.3.1.1 Data collection

The wireless network at our campus consists of about 500 access points (routers) located in 180 buildings, which also cover some of the inter-building areas. The network requires the users to authenticate with unique credentials. The routers store the anonymized list of connected users every 10 minutes along with a timestamp. It is thus possible to track the changing location of each individual in an anonymous fashion; whenever a user u moves from a location l_w, such as a lecture hall to location l_z, for example a cafeteria, this is recorded as presence of the user u at location l_w in time t_w followed by presence of the user in locations l_x,l_y which are located between l_w and l_z and finally by occurrence of this user at location l_z in time t_z. Due to the fact that we use credentials to identify users and not MAC addresses of devices, the analysis is not prone to errors originating from one person using several different devices. The location of faculty members and other university workers (like the night watchmen) has not been discarded to avoid jeopardizing the privacy and potentially security of the campus, when the dataset is made public.

4.3.1.2 Summary statistics

The dataset grows by about 1.5 million data points per month during the semester, and the analysis in this paper is focused on the data collected between end of August and end of December 2011 – about 6 million data points. 7700 unique users were registered in that period. See Figure4.1 for the distribution of their activity.

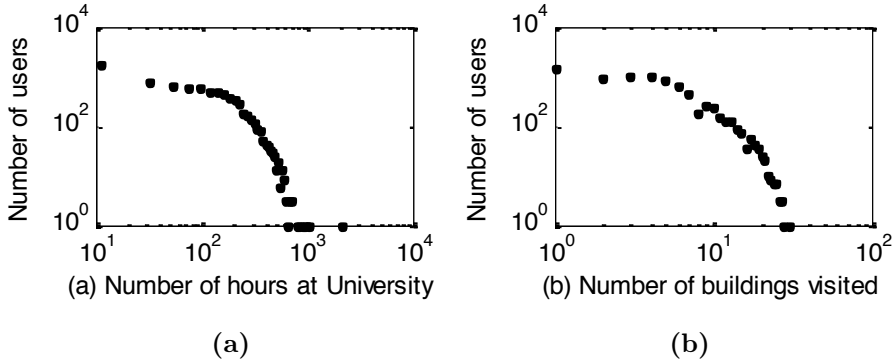


Figure 4.1: Histograms of number of hours spent at the campus per user (a), and number of buildings each of them visited (b)

4.3.1.3 Preprocessing

As the first preprocessing step, the data is discretized into time bins each spanning 20 minutes, which is twice the sampling period of the routers. If a smaller value was chosen, some of the actual co-occurrences might have failed to be recorded. On the other hand, by increasing the span of the time bin, we would assume co-occurrences between people who visited a particular location in different moments, without the possibility of actually meeting each other. For each time bin, the total number of people present in each access point and each building is calculated. Then, sparse binary tables are created to mark the location of each person who delivered the self-report at each time bin with respect to access points and buildings.

4.3.1.4 Challenges

Our method of data collection allows us to trace thousands of people without their active participation but it also sets a number of challenges. The main limitation is that a person must be connected to the campus Wi-Fi network using a laptop computer, a smartphone, or a tablet. This requirement creates many scenarios, which result in missing data. The most important are:

- no data points are collected outside of campus and not all campus is equally well covered by the network
- many students use the computers available on campus which are connected to wired network instead of their personal computers and Wi-Fi

- when people work in groups, not all the members need to use their computers
- the default settings on most smartphones are that the Wi-Fi radio turns off when the screen is off; in normal usage the screen is off most of the time
- due to the reception problems, many students turn the Wi-Fi radio off at all times in their smartphones and only rely on 3G for data transfers
- most students do not use computers during lunch time

Furthermore, the data is also affected by noise due to the following factors:

- people leave their laptops on overnight or for several days without actually being close to their machines
- some Wi-Fi access points are moved around the campus and their location is only updated once per day
- some routers are not mapped to any particular location

Finally, the main challenge is that work-related meetings are much more likely to result in recorded co-occurrences between people than encounters which are fun-related and thus with stronger social signal. For example, it is more probable that two people who meet to work on a project would use their computers, than if they just meet to have a beer together.

4.3.2 Self-report

4.3.2.1 Level of association

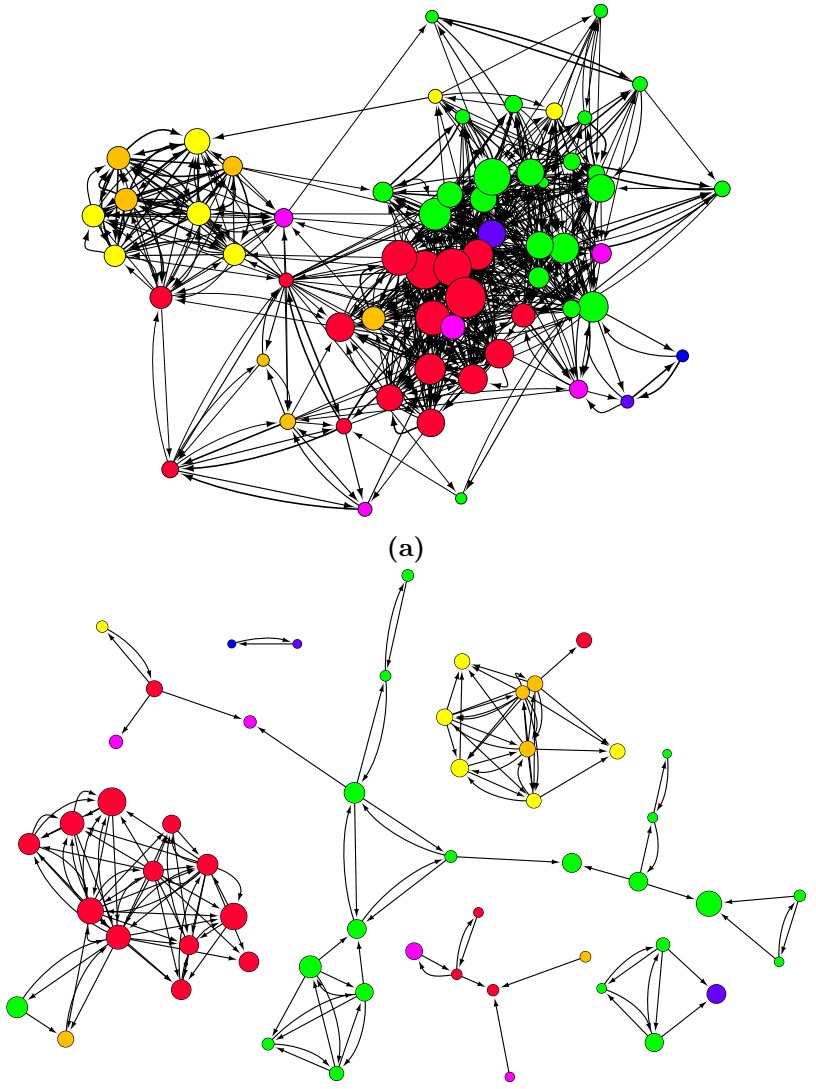
Participants of two university classes were asked to assign grades to others based on how well they knew them. Student numbers of the respondents were anonymized with the same encryption as the Wi-Fi scan results, so that the self-reports could be compared to the behavioral data. Note, that the survey was only conducted once, so it was not possible to observe the temporal evolution of the network. The students were presented with the name, surname and a picture of each other participant and the following description of association grades: 1 — "I do not know her/him at all", 2 — "I know who she/he is, but not much more", 3 — "I know who she/he is, we might have taken some courses

together", 4 — "I see her/him around, we might have been in a project group", 5 — "She/he is my good friend". In both courses around 60 persons responded. Their answers covered not only other respondents, but also some of the students, who did not volunteer for the survey. In order to avoid privacy concerns these grades were discarded. As a result, two square matrices with association grades were created, one for each course. See Figure 4.2a and Figure 4.3a for the visualizations of full social networks as reported by subjects. Note that most of the dyadic links occur between people who have started the university in the same year (their nodes have the same color, see Table 4.1). This effect was even more clearly visible in the networks where edges exist only whenever the highest grade was given (5 — "good friends"), as shown in Figure 4.2b and Figure 4.3b. It also occurs that the friendship is not always symmetrical – in class A 56 out of 160 links reported as 5's are not reciprocated (35%); in class the respective values are 22 out of 58 (38%). Due to its duration of only 3 weeks, Class B occurred much less interconnected, and therefore is not analyzed further.

4.3.2.2 Formation and reciprocity of friendship with respect to ranking

Ball and Newman[BN12] in their recent paper state that the friendship reciprocity is correlated with ranking of each individual, i.e. most of the reciprocated friendships are bound between persons with the same ranking. Most non-reciprocal links, on the other hand, originate from lower-ranked persons who described higher-ranked peers as friends. Ball and Newman define ranking as a function positively correlated with the year of commencing the school and total number of friends. Here, we only look into the year, as the number of friends is to be inferred from the behavioral data.

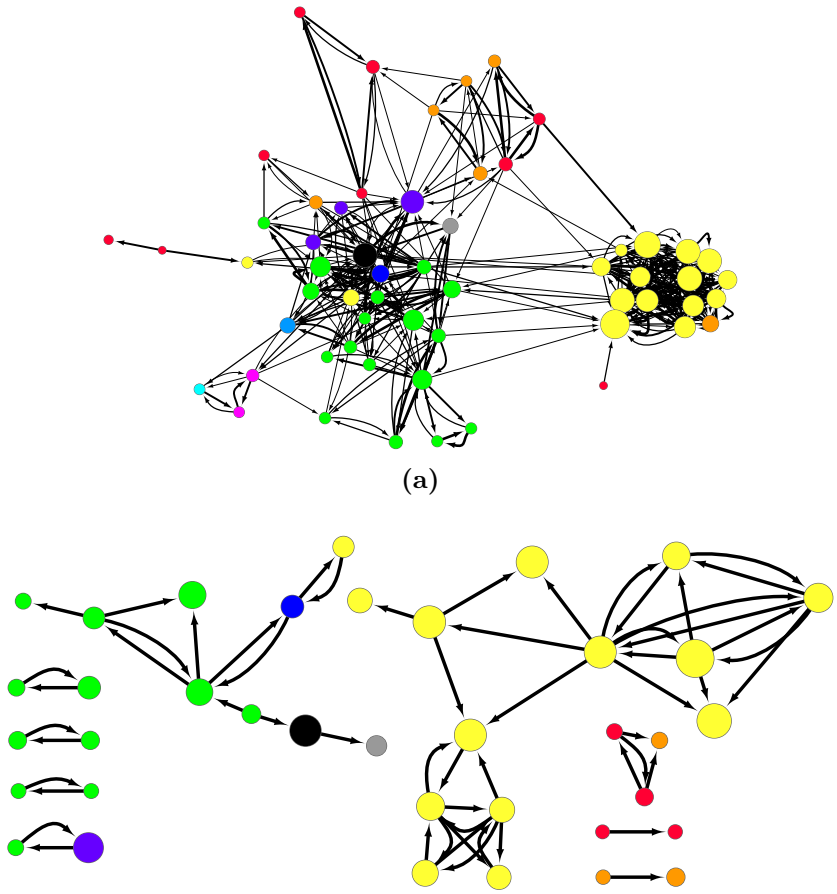
Figure 4.4a confirms that the highest probability of friendship in class A occurs between two persons who started the university in the same year. Moreover, the regression fit in Figure 4.4b shows that users with higher ranking are less likely to reciprocate the friendships originating from lower-ranked peers. In class B, the same-rank students are still most likely to create friendships, but the reciprocation probability does not decay with the growing rank difference.



(a)

(b)

Figure 4.2: Social network in class A according to the self-reports. Edge width symbolizes the grade, node color denotes the starting year at the university, node size denotes the sum of received grades (without 1's). Two views are presented: full network (a), and highest-grade friendships only (b).



(a)

(b)

Figure 4.3: Social network in class B according to the self-reports. Edge width symbolizes the grade, node color denotes the starting year at the university, node size denotes the sum of received grades (without 1's). Two views are presented: full network (a), and highest-grade friendships only (b).

	'91		'01		'02		'03		'05		'06
	'07		'08		'09		'10		'11		

Table 4.1: Explanation for color-coded first year at the university

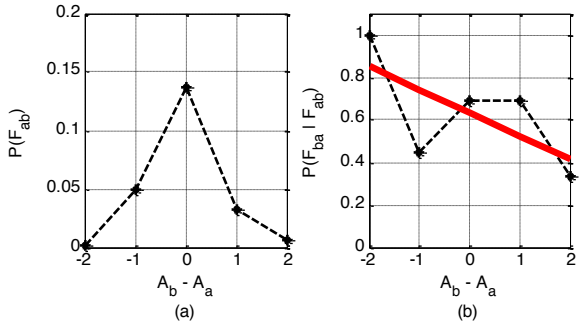


Figure 4.4: (a) Probability of friendship link between students a and b in class A as a function of rank difference; (b) Probability of friendship reciprocity as a function of rank difference; The red solid line represents the linear regression fit; Symbols used in the plots: F_{ab} -– User a declares friendship with user b, A_a - ranking of user a. Age differences of more than 2 years are represented as 2 years differences.

4.3.3 Correlation between behavioral and self-reported data

Before we commence with the actual modeling it is necessary to verify whether the self-reported data is correlated with the co-location data. Findings from this section will be used as a benchmark for examining the performance of models used in further analysis.

Firstly, we examine the correlation between the number of co-occurrences and the self-reported grade, see Table 4.2. Columns A and B contain the correlation values (ρ) for the two disjoined classes of participants.

The values listed in the Table 4.2 show that even though the number of co-occurrences and the self-reported grade are highly correlated, using just the raw count of co-location does not explain the whole grading. In further analysis we binarize the links as 1 — grade 5 and 0 — grade lower than 5. The most straight-forward unsupervised way for predicting the friendship links in the data is by setting a threshold on the number of co-occurrences to then positively "predict" dyads with higher number as friends and dyads with lower number as non-friends. By moving the threshold from the highest to the lowest value, we obtain the Receiver Operating Characteristic (ROC), which is a tool widely used for evaluation of prediction performance and especially suitable for datasets with severe class imbalance [TSK06]. The curves are presented in Figure 4.5 for class A and Figure 4.6 for class B. ROC depicts the trade-off between true positive rate (TPR — fraction of correctly classified positives among all positives in the dataset) and false positive rate (FPR – fraction of negatives wrongly classified as positives among all negatives in the dataset). The area under the ROC curve (AUC) is adopted as a measure of performance. An ideal model achieves $TPR = 1$, $FPR = 0$ and thus $AUC = 1$, while the results from a random classifier fall on the diagonal of the plot yielding $AUC = 0.5$. In unbalanced datasets the ROC can be partially misleading as even low FPR produces a lot of false positives. Therefore, yet another value is introduced and plotted against TPR – the positive predictive value (PPV). PPV shows the fraction of observations classified as positive, which are in fact positive.

By comparing the performance between the two classes, we can see that al-

	All grades		5's only	
	A	B	A	B
ρ	0.5551*	0.5292*	0.4360*	0.3946*

Table 4.2: Correlation between the number of co-occurrences between two persons and the self-reported grade. *: $p_{val} < 10^{-100}$

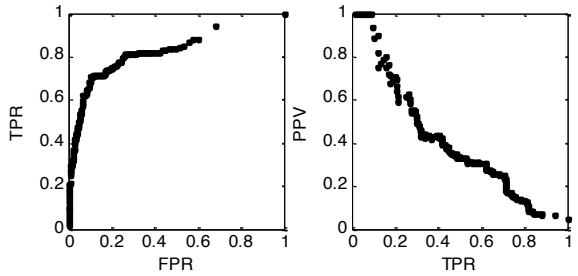


Figure 4.5: Receiver Operator Curve (on the left) for thresholding number of co-occurrences between participants in class A yielding $AUC = 0.835$ and PPV as a function of TPR (on the right). The results can be interpreted as follows: the model can recall 10% of friends ($TPR = 0.1$) without misclassifying any non-friends ($PPV = 1$, $FPR = 0$); by lowering the threshold, 30% of friendships can be recalled ($TPR = 0.3$), but only 50% of all links classified as positive are actually friends ($PPV = 0.5$, $FPR = 0.01$). Furthermore, it can be seen that 6% of friendship links have no support in the data

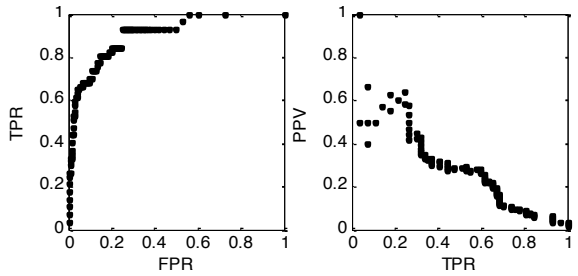


Figure 4.6: ROC for thresholding number of co-occurrences between participants in class B yielding $AUC = 0.909$ and PPV as a function of TPR (on the right). Interpretation: only two links can be recalled without false positives ($TPR = 0.035$); by decreasing the threshold, 26% of links can be recalled, but only 50% of all positively classified observations are actually friends links. Furthermore, it can be seen that all friendships have support in the data. By setting the threshold low enough to recall all friendships, the model yields PPV of only 0.031 and FPR of 0.56

	NF_{out}		NF_{in}	
	A	B	A	B
t_S	0.2787*	0.4313***	0.3078*	0.2374
t_S/t_T	0.2676*	0.4122**	0.3613**	0.5284***

Table 4.3: Correlation between the number of good friends a person reported (NF_{out}), number of peers who grade a person as a good friend (NF_{in}), the time that person spends with other people at campus (t_S) and the fraction of the time that person spends with others with respect to the total time she spends on campus (t_S/t_T). *: $p_{val} < 0.05$, **: $p_{val} < 0.01$, ***: $p_{val} < 0.001$

though thresholding raw co-occurrences works better in class A, the AUC is lower than for class B. This shows that the AUC value can only be used to compare performance of different models on the same dataset and not the same model across datasets. Therefore, for each model under scrutiny we also report the maximum TPR yielded with PPV equal to 50% (the higher the better) as well as the highest PPV yielded with $TPR = 0.5$ (the higher the better). These metrics will be referred to as $TPR@PPV_{0.5}$ and $PPV@TPR_{0.5}$.

Next, we examine the tendency of persons to describe others as friends and be perceived as ones. Following symbols are used in Table 4 to present the influential factors: t_T -– total time spent on campus by a person, t_S – social time, spent with any number of other respondents, t_S/t_T - proportion of total time which is spent with other respondents, NF_{out} – number of people who a person regards as good friends, NF_{in} – number of people who regard a person as a good friend. Note, that the correlations with t_T are not listed as none is of statistical significance ($p_{val} > 0.05$). For both classes the correlation between the number of people who regard a particular person as a friend is highly correlated with the ratio of social time with regard to total time spent at the university, and this correlation is statistically significant.

4.4 Modeling

In this section we describe the machinery used for link prediction as well as the features, which are extracted from spatio-temporal traces to aid the prediction. Even though we showed that the year of starting the university could be a useful feature in friendship link prediction, we choose to concentrate on behavioral features, which do not require prior knowledge of the observed population.

Modeling is performed on two datasets to answer two questions:

- how to recognize friends from non friends,
- how to recognize reciprocal friends from non-reciprocal friends.

We first analyze the utility of single features and then use more advance machine learning approaches, which combine multiple features, in order to increase the predictive performance beyond the levels offered by single attributes. Modeling is performed using algorithms implemented in *scikit-learn* [PVG+11] module for Python. The *scikit-learn* supports outputting probability estimations instead of final results of classification, it is thus possible to calculate the area under ROC as well as other metrics used in this paper to evaluate the performance of each classifier. Optimal parameters of each classifier are estimated through grid search with cross-validation using train and test sets, which are created randomly but preserving the prior probabilities of classes.

4.4.1 Ensemble learning

We use a number of machine learning approaches, which combine multiple fitted models to arrive at a better final prediction than any single of the incorporated models [TSK06]. Such approach is called *ensemble learning*. Two types of ensemble learning methods are used in this paper: *bagging* and *boosting*.

4.4.1.1 Bagging

In *bagging* each constituent model is fitted on a random subsample of training data and has a vote of equal importance while predicting the class of new samples. Because the training of each model is independent of other models, this approach can easily be parallelized. *Random Forest* is the bagging approach used in this paper.

4.4.1.2 Boosting

Initially, each sample of data has equal importance in *boosting*; models are trained sequentially and samples which are mis-classified have their weight increased and are thus more frequent in training sets for subsequent models. This

approach is more susceptible to over-fitting than bagging and cannot be parallelized. *AdaBoost* and *Gradient Tree Boosting* are the boosting approaches used in this paper.

4.4.2 Prediction features

The machine learning algorithms are fed with multi-variate representation of the dataset, where each variable is a feature extracted from the raw dataset. For each feature presented, there is a histogram of distribution of that feature's values for friends and non-friends dyads, see Figure 4.8, as well as for non-reciprocal and reciprocal friends, see Figure 4.9. Note, that the distributions are very different: in all features, most of non-friends dyads have a value of zero and the distributions are thin-tailed. Friends dyads peak at higher values and have thicker-tailed distributions. Even though at first glimpse it might seem that some of the single features provide near-perfect separation between these two classes of dyads, it is not the case, because of the class imbalance problem. For example, in Figure 4.8l, more than 40% of friends dyads and only 20% of non-friends dyads have entropy of their time of the week meetings higher than 1. However, If we just assume that dyads with entropy higher than 1 are friends, we classify 61 actual friends dyads correctly and 315 non-friends dyads as friends, thus arriving at very low precision.

4.4.2.1 Raw co-occurrences

As described in the previous section, we use the raw co-occurrence count between persons as a benchmark for evaluating the predictive performance of other features. Feature ID: `coocs_raw`, distributions presented in Figure 4.8a.

4.4.2.2 Co-occurrences weighted with respect to buildings

While other researchers use entropy to weight the social impact of meetings, our data allows us to introduce a more precise measure. We use anonymous statistics to estimate the number of all people present in the building in each time bin. We assume the social importance of each co-occurrence to be inversely proportional to the number of people – if only a few persons are there in a location, it is more probable that there is a social bond between them compared to the situation when dozens of people are present. Feature ID: `coocs_w`, see Figure 4.8b.

4.4.2.3 Timely arrival and leaving

We propose that if two persons arrive at a location at the same time and/or leave the location synchronously it yields a stronger signal than if two people are in the same location, but their arrival and leaving are not synchronized. The value is weighted by the number of people who arrived and/or left the building in each particular time bin. Thus, timed arrival of many people in the beginning of the scheduled classes is not as strong a signal as synchronized arrival of a few persons an hour before the class begins. Feature ID: `arr_leav`, see Figure 4.8c.

4.4.2.4 Time intervals between meetings

As shown in Figure 4.7, friends and non-friends show distinctly different patterns with regard to time interval between co-occurrences. Non-friends are very likely to meet in weekly-intervals while friends tend to meet more than once per day and several times per week. 30% of all co-occurrences between friends happen 24 hours from last co-occurrence and 75% within 6 days. These values are respectively 13% and 49% for non-friends. We derive the following features to exploit the differences in co-occurrence patterns:

- number of separate meetings (ID: `t_n`, Figure 4.8d)
- percentage of meetings which happened within 24 hours from previous meetings (ID: `t_1st_day`, Figure 4.8e)
- percentage of meetings which happened within 6 days from previous meetings (ID: `t_6_days`, Figure 4.8f)

Preliminary analysis showed that fraction of meetings which happened with n-week intervals, mean interval between meetings, and variance in intervals between meetings are not significantly correlated with the friendship and are therefore not used for further analysis.

4.4.2.5 Common travel

We propose distance travelled together as a measure of association. Whenever two persons co-occur in different locations in subsequent time bins, we calculate the geographic distance between the buildings based on their known coordinates. Distances traveled together are then summed for each dyad. The following

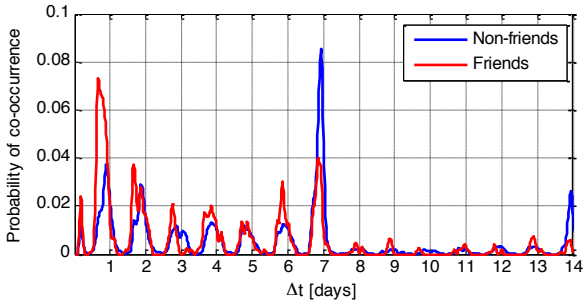


Figure 4.7: Probability of interval between meetings for friends and non-friends. For readability only the first two weeks are presented and the results were convoluted with a Hamming window spanning 4 hours. Intervals equal or lower than 1 hour were treated as continuous meetings and thus discarded from the time differences.

features are derived: total distance (ID: `tr_dist`, Figure 4.8g), number of travels (ID: `tr_n`, Figure 4.8h) and mean distance traveled (ID: `tr_mean`, Figure 4.8i).

4.4.2.6 Correlation between time vectors

Due to limited number of possible locations, the activity of each user can be described as a set of binary vectors of presence at each location. Each dyad of users can then be assigned a similarity score by calculating the correlation between their activity vectors for each building and then summing all the statistically significant correlations (ID: `corr`, Figure 4.8j)

4.4.2.7 Spatial entropy of co-occurrences

We introduce entropy as a measure of diversity of locations where two persons meet (ID: `entr_loc`, Figure 4.8k). The intuition is that if two persons are seen together in different locations (high entropy) then it is more likely that these persons are friends compared to a dyad, which only meets in one building (low entropy).

Let C_k be the number of all co-occurrences of dyad d_k, c_ik be the number of co-occurrences of this dyad in location l_i and ϕ_k the set of all locations where the dyad d_k met. Then $q_{ik} = c_{ik}/C_k$ is the fraction of meetings in location l_i with respect to C_k and $q_{1k}, ..., q_{Nk}$ is the discrete probability distribution which

denotes how likely the dyad d_k co-occurred in each particular location. Then, the dyadic entropy E_k is defined in Equation 4.1

$$E_k = -\sum_{p_i \in \phi_k} q_{ik} \log q_{ik} \tag{4.1}$$

This metric has been used in other works to describe the location-specific diversity [CTH+10, SNM11] and weekly schedules of individual participants [EP06b, CTH+10].

4.4.2.8 Temporal entropy of co-occurrences

We follow similar intuition to introduce entropy as a measure diversity of times when people meet. The features include:

- entropy of hour in week (ID: `entr_w_h`, Figure 4.8l), entropy of time of day in the week (ID: `entr_w_t`, Figure 4.8m): people who meet in various time slots across weeks are more likely to be friends than those who only meet in specific moments (for example just in the classes)
- entropy of day of week (ID: `entr_w_d`, Figure 4.8n): as shown in Figure 4.7 non-friends tend to meet on a specific day in the week. Thus, their dyads have lower entropy than friends who are more likely to meet several times a week
- entropy of the week number (ID: `entr_w_id`, Figure 4.8o): people who met in different weeks across the semester have higher entropy than those who only met during a part of the semester

4.4.2.9 Specificity

The asymmetric specificity S_{ij} defined as fraction of time person p_i spends with person p_j with respect to the total time spent on campus by person p_j. As shown in Table 4.3, the fraction of social time with respect to total time is more indicative of being perceived as a friend than only the social time (ID: spec, Figure 4.8p).

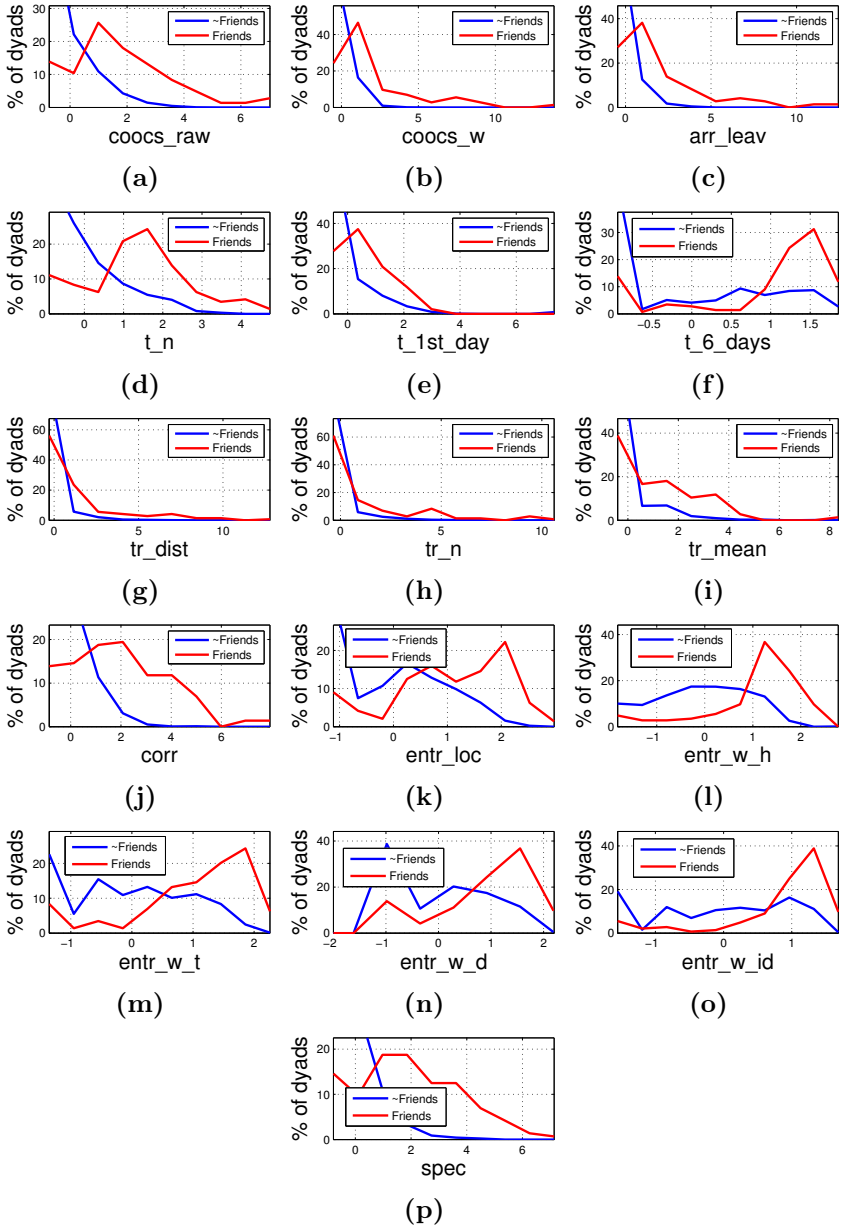


Figure 4.8: Histograms of non-friends and friends dyads' distributions of each prediction feature. In all features, most of non-friends dyads have a value of zero and the distributions are thin-tailed. Friends dyads peak at higher values and have thicker tailed distributions.

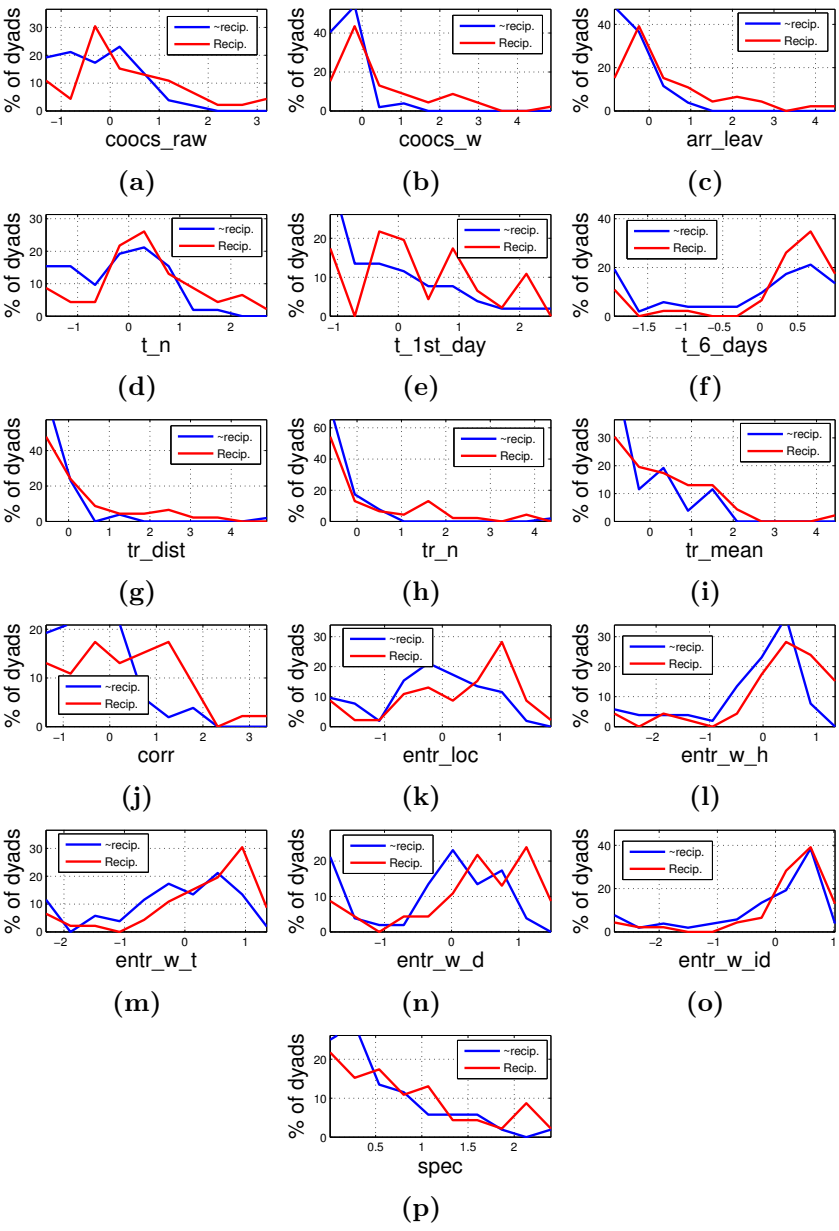


Figure 4.9: Histograms of reciprocal and non-reciprocal friends dyads' distribution of each prediction feature. Unlike in the friends/non friends problem, the "negative" class (here: non-reciprocal friends) is not concentrated around the minimum value of the feature, and the distributions of the two classes are much more similar.

4.5 Results

In this section we first evaluate the predictive performance of each feature in an unsupervised thresholding task using the AUC as well as TPR with PPV equal to 50%. Next, we train several supervised learning models and elaborate on the results from the best performing approach.

4.5.1 Performance of single features

As shown in Figure 4.10, the performance of thresholding raw co-occurrences performs better than most of the derived features. It is also clear that for both datasets there are four features which perform better than the benchmark: weighted co-occurrences, synchronized arrival and leaving, correlation between activity vectors, and specificity as shown in Figure 4.10.

4.5.2 Supervised learning performance

In this section we evaluate predictive performance of three models: AdaBoost, Gradient Tree Boosting and Random Forest. AUC, $PPV@TPR_{0.5}$, $TPR@PPV_{0.5}$ are used for comparative evaluation, box plots for these values are provided to qualitatively inspect stability of these models. Additionally, importance of features is presented to aid the interpretation of results.

Figure 4.11a shows the AUC score for the three ensemble methods for the Friends vs Non-friends classification problem. Random Forest approach outperforms two other methods, with significantly higher median value as well as less variability in perfomance across different fits. Gradient Boosting offeres a slight improvement over AdaBoost, and all the classifiers perfom significantly better than a random classifier. Figures 4.13a 4.13b 4.13c present the experimental importance of each feature while performing the classification. Specificity is clearly the most important feature in all the models, but correlation and weighted cooccurrences are also used.

Figure 4.11b shows the AUC score for the three ensemble methods for the Reciprocal vs. Non-reciprocal friends classification problem. Here, the problem is clearly more difficult - the classifiers still perform better than random on average, but some fits do not. Also, there is no single classifier that works much better than the others. Figures 4.13d 4.13e 4.13f present the experimental importance of each feature while performing the classification. Random forest

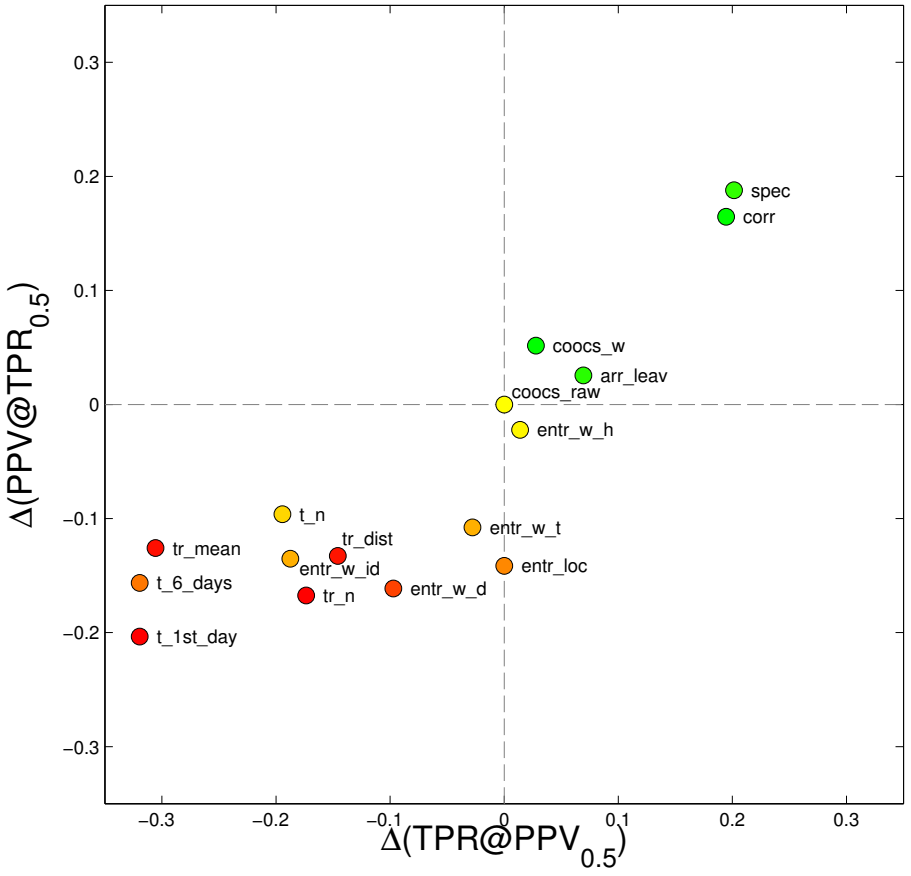


Figure 4.10: Performance evaluation of thresholding single features for the two classes of students. The color denotes the change in AUC relative to performance of raw co-occurrences count -– green is the increase, red -– decrease, and yellow symbolizes no change. In the center of the plot there is the `coocs_raw` used as the benchmark. Features which yield better results in $PPV@TPR_{0.5}$ and $TPR@PPV_{0.5}$ are in the first quadrant, while features with worse performance are in the third quadrant

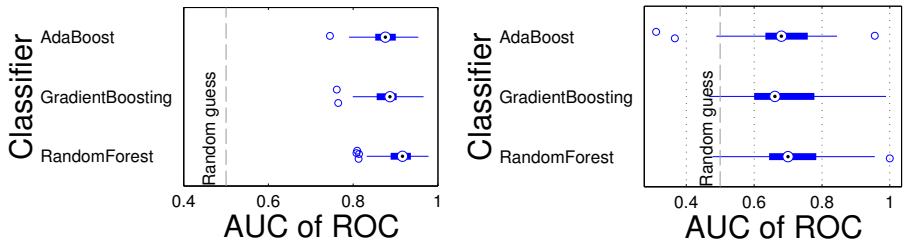


Figure 4.11: Box plots of *AUC* of *ROC* for three ensemble methods and two classification problems. (a) In friends vs non-friends problem, Random Forest classifier clearly outperforms two other approaches, with higher median *AUC* and less variance across fits. All three models perform much better than random. (b) In reciprocal vs non-reciprocal problem, all models have fits not performing better than random, but Random Forest still performs slightly better than the other two

bases its classification in this problem on weighted co-occurrences and timed arrival with higher weight than on specificity. The two other classifiers, however still give specificity the highest importance. Since the performace of the three classifiers is similar in this problem, strong conculsion about what makes friends reciprical cannot be found.

Finally, Figure 4.14 visualizes the utility of employing sophisticated machine learning approaches over basing predictions on single features. All three classifiers perfom much better in terms of all analyzed metrics, with Random Forest outperforming other methods.

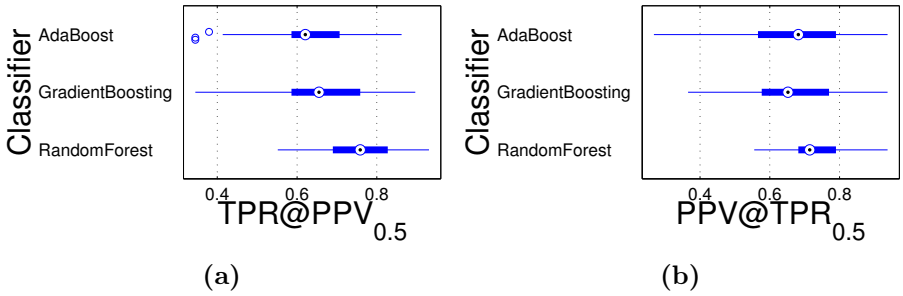


Figure 4.12: Box plots of (a) $TPR@PPV_{0.5}$ and (b) $PPV@TPR_{0.5}$ for three ensemble methods and the Friends vs Non-friends problem. Random Forest classifier again shows highest performance in both features and least variance across fits.

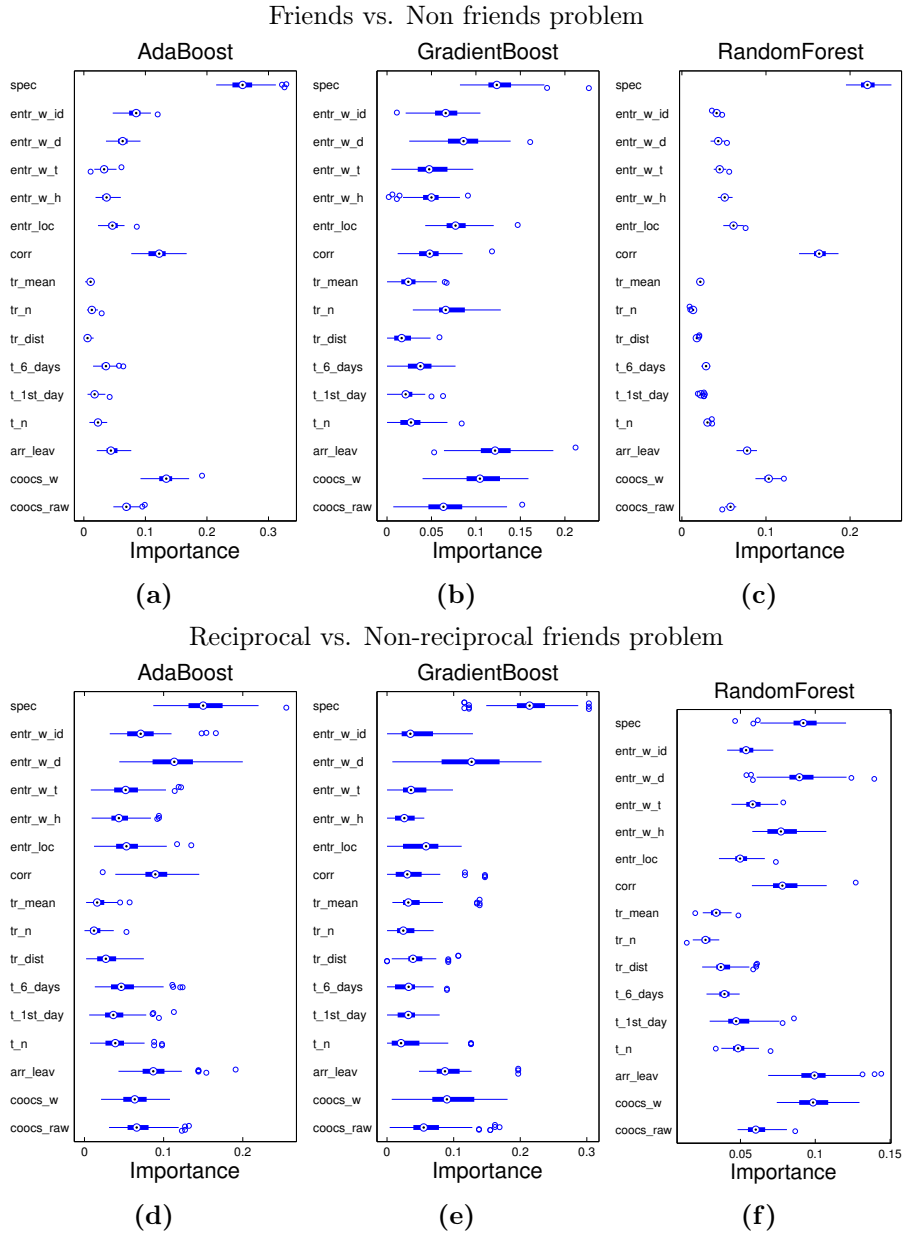


Figure 4.13: Visualization of importance of features among different classifiers for two problems. Note, that the best-perfoming classifier (Random Forest) uses different features to solve the two problems.

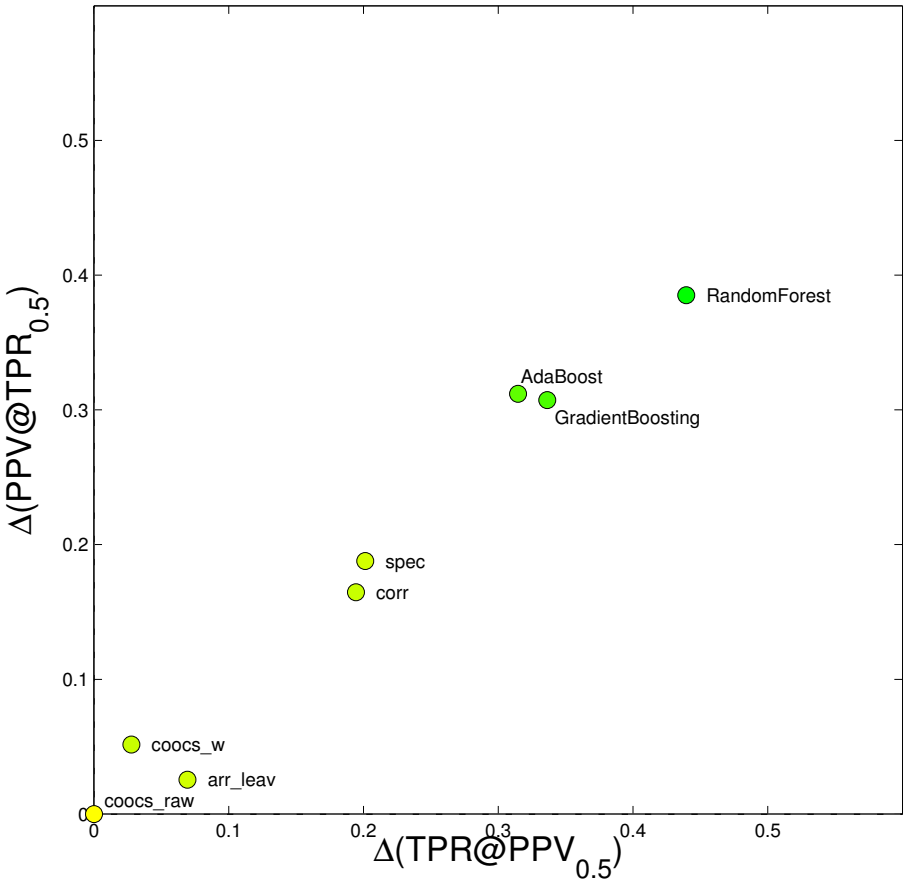


Figure 4.14: Predictive performance of single features compared with the performance of more sophisticated models - there is no single observable behavioral pattern among those proposed, which would indicate friendship as accurately as more complex models based on sets of behaviors.

4.6 Discussion

In this paper, we have analyzed a real-world social network of students of one class at a European university. We found distinct behavioral features, which enabled us to accurately find friendship links in a highly imbalanced dataset, as well as understand differences in co-occurence patterns between reciprocal and non-reciprocal friends.

However, it is crucial to remember that we only analyzed a small group based in a very specific context. The students, who provided their self-report, attend the same class and are thus "forced" to co-occur quite often. There is therefore much more dyadic interaction between them than in a set of random students, who do not study together, let alone a set of random people in, for example, a city. It is clear that the method is not flawless, as people have social networks spanning across universities, cities, countries and continents, and what can be observed from a location specific WiFi network is only a small fraction of peoples' social life.

The analysis would benefit greatly from data of higher quality and originating from a bigger number of independent sensors, as some of the respondents were not even once connected to the system WiFi during the study. Moreover, more respondents, spanning multiple courses, and featuring a wider range of academic status would make the findings more generizable. Finally, periodic self-reports would enable temporal analysis of the network dynamics.

4.7 Conclusions

Due to the effective range of ~100 m WiFi Access Points are susceptible to falsely sense many interactions, which did not take place. On the other hand, many actual face-to-face interactions will go unnoticed, if at least one person is not connected to the wireless network. However, as we have shown on real-world data, it is possible to recover meaningful signal from this noisy data. This paper shows the pipeline of social discovery, starting from raw WiFi router logs and finishing by reconstructing cognitive social network.

We presented a number of features, which, by putting each co-presence into a spatio-temporal context, enabled deeper understanding of the nature of interactions between strangers, as well as between reciprocal and non-reciprocal friends. While raw count of co-presence events is usually used to measure interaction, we showed that certain publicly observable patterns of interaction are

much more significant to participants, a notion which is not captured by the raw count.

Finally, using state of the art machine learning approaches, we were able to reconstruct cognitive social network of the experiment's participants. In future research, we will employ the derived behavioral patterns and mathematical approaches to analyze social networks measured using not only fixed-location, but also mobile sensors.

Appendix A

Measuring Personalization of Web Search

Measuring Personalization of Web Search

Aniko Hannak
Northeastern University
ancsaaa@ccs.neu.edu

Piotr Sapieżyński
Technical University of Denmark
sapiezynski@gmail.com

Arash Molavi Kakhki
Northeastern University
arash@ccs.neu.edu

Balachander Krishnamurthy
AT&T Labs–Research
bala@research.att.com

David Lazer
Northeastern University
d.lazer@neu.edu

Alan Mislove
Northeastern University
amislove@ccs.neu.edu

Christo Wilson
Northeastern University
cbw@ccs.neu.edu

ABSTRACT

Web search is an integral part of our daily lives. Recently, there has been a trend of personalization in Web search, where different users receive different results for the same search query. The increasing personalization is leading to concerns about *Filter Bubble* effects, where certain users are simply unable to access information that the search engines' algorithm decides is irrelevant. Despite these concerns, there has been little quantification of the extent of personalization in Web search today, or the user attributes that cause it.

In light of this situation, we make three contributions. First, we develop a methodology for measuring personalization in Web search results. While conceptually simple, there are numerous details that our methodology must handle in order to accurately attribute differences in search results to personalization. Second, we apply our methodology to 200 users on Google Web Search; we find that, on average, 11.7% of results show differences due to personalization, but that this varies widely by search query and by result ranking. Third, we investigate the causes of personalization on Google Web Search. Surprisingly, we only find measurable personalization as a result of searching with a logged in account and the IP address of the searching user. Our results are a first step towards understanding the extent and effects of personalization on Web search engines today.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*

Keywords

Personalization; Web search; Measurement

1. INTRODUCTION

Web search services like Bing and Google Web Search (Google Search) are an integral part of our daily lives; Google Search alone receives 17 billion queries per month from U.S. users [52]. People use Web search for a number of reasons, including finding authoritative sources on a topic, keeping abreast of breaking news, and making purchasing decisions. The search results that are returned, and their order, have significant implications: ranking certain results higher or lower can dramatically affect business outcomes (e.g., the popularity of search engine optimization services), political elections (e.g., U.S. Senator Rick Santorum's battle with Google [18]), and foreign affairs (e.g., Google's ongoing conflict with Chinese Web censors [46]).

Recently, major search engines have implemented *personalization*, where different users searching for the same terms may observe different results [1, 34]. For example, users searching for "pizza" in New York and in Boston may receive different, but locally relevant restaurant results. Personalization provides obvious benefits to users, including disambiguation and retrieval of locally relevant results.

However, personalization of Web search has led to growing concerns over the *Filter Bubble* effect [9], where users are only given results that the personalization algorithm thinks they want (while other, potentially important, results remain hidden). For example, Eli Pariser demonstrated that during the recent Egyptian revolution, different users searching for "Tahrir Square" received either links to news reports of protests, or links to travel agencies [26]. The Filter Bubble effect is exacerbated by the dual issues that most users do not know that search results are personalized, yet users tend to place blind faith in the quality of search results [25].

Concerns about the Filter Bubble effects are now appearing in the popular press [35, 38], driving growth in the popularity of alternative search engines that do not personalize results (e.g., `duckduckgo.com`). Unfortunately, to date, there has been little scientific quantification of the basis and extent of search personalization in practice.

In this paper, we make three contributions towards remedying this situation. *First*, we develop a methodology for measuring personalization in Web search results. Measuring personalization is conceptually simple: one can run multiple searches for the same queries and compare the results. However, accurately attributing differences in returned search results to personalization requires accounting for a number of phenomena, including temporal changes in the search index, consistency issues in distributed search indices, and A/B tests being run by the search provider. We develop a methodology that is able to control for these phenomena and create a command line-based implementation that we make available to the research community.

WWW 2013, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2035-1/13/05.

Second, we use this methodology to measure the extent of personalization on Google Web Search today. We recruit 200 users with active Google accounts from Amazon's Mechanical Turk to run a list of Web searches, and we measure the differences in search results that they are given. We control for differences in time, location, distributed infrastructure, and noise, allowing us to attribute any differences observed to personalization. Although our results are only a lower bound, we observe significant personalization: on average, 11.7% of search results show differences due to personalization, with higher probabilities for results towards the bottom. We see the highest personalization for queries related to political issues, news, and local businesses.

Third, we investigate the causes of personalization, covering user-provided profile information, Web browser and operating system choice, search history, search-result-click history, and browsing history. We create numerous Google accounts and assign each a set of unique behaviors. We develop a standard list of 120 search queries that cover a variety of topics pulled from Google Zeitgeist [14] and WebMD [48]. We then measure the differences in results that are returned for this list of searches. Overall, we find that while the level of personalization is significant, there are very few user properties that lead to personalization. Contrary to our expectations, we find that only being logged in to Google and the location (IP address) of the user's machine result in measurable personalization. All other attributes do not result in level of personalization beyond the baseline noise level.

We view our work as a first step towards measuring and addressing the increasing level of personalization on the Web today. All Web search engines periodically introduce new techniques, thus any particular findings about the level and causes of personalization may only be accurate for a small time window. However, our methodology can be applied periodically to determine if search services have changed. Additionally, although we focus on Google Search in this paper, our methodology naturally generalizes to other search services as well (e.g., Bing, Google News).

2. BACKGROUND

We now provide background on Google Search and overview the terminology used in the remainder of the paper.

2.1 A Brief History of Google

Personalization on Google Search. Google first introduced "Personalized Search" in 2004 [17], and merged this product into Google Search in 2005 [1]. In 2009, Google began personalizing search results for all users, even those without Google accounts [15]. Recently, Google started including personalized content from the Google+ social network into search results [33]. For example, users may see Web pages which were shared or "+1'd" by people in their Google+ circles alongside normal Google search results.

There is very little concrete information about how Google personalizes search results. A 2011 post on the official Google blog states that Google Search personalizes results based on the user's language, geolocation, history of search queries, and their Google+ social connections [32]. However, the specific uses of search history data are unclear: the blog post suggests that the temporal order of searches matters, as well as whether users click on results. Similarly, the specific uses of social data from Google+ are unknown.

Figure 1: Example page of Google Search results.

Google Accounts. As the number and scope of the services provided by Google grew, Google began unifying their account management architecture. Today, Google Accounts are the single point of login for all Google services. Once a user logs in to one of these services, they are effectively logged in to all services. A tracking cookie enables all of Google's services to uniquely identify each logged in user. As of May 2012, Google's privacy policy allows between-service information sharing across all Google services [45].

Advertising and User Tracking. Google is capable of tracking users as they browse the Web due to their large advertising networks. Roesner et al. provide an excellent overview of how Google can use cookies from DoubleClick and Google Analytics, as well as widgets from YouTube and Google+ to track users' browsing habits [31].

2.2 Terminology

In this study, we use a specific set of terms when referring to Google Search. Each *query* to Google Search is composed of one or more keywords. In response to a query, Google Search returns a *page* of *results*. Figure 1 shows a truncated example page of Google Search results for the query "coughs." Each page contains ≈10 results (in some cases there may be more or less). We highlight three results with red boxes in Figure 1. Most results contain ≥ 1 links. In this study, we only focus on the *primary link* in each result, which we highlight with red arrows in Figure 1.

In most cases, the primary link is *organic*, i.e., it points to a third-party website. The WebMD result in Figure 1 falls into this category. However, the primary link may point to another Google service. For example, in Figure 1 the "News for coughs" link directs to Google News.

A few services inserted in Google Search results do not include a primary link. The Related Searches result in Figure 1 falls into this category. Another example is Google Dictionary, which displays the definition of a search keyword. In these cases, we treat the primary link of the result as a descriptive, static string, e.g., "Related" or "Dictionary."

3. METHODOLOGY

In this section, we describe our experimental methodology. First, we give the high-level intuition that guides the design of our experiments, and identify sources of noise that can

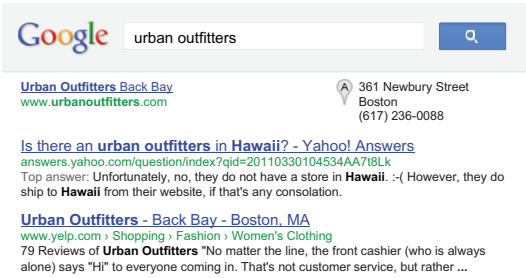


Figure 2: Example of result carry-over, searching for "hawaii" then searching for "urban outfitters."

lead to errors in data collection. Second, we describe the implementation of our experiments. Lastly, we introduce the queries we use to test for personalization.

3.1 Experiment Design

Our study seeks to answer two broad questions. First, *what user features influence Google's search personalization algorithms?* This question is fundamental: outside of Google, nobody knows the specifics of how personalization works. Second, *to what extent does search personalization actually affect search results?* Although it is known that Google personalizes search results, it is not clear how much these algorithms actually alter the results. If the delta between "normal" and "personalized" results is small, then concerns over the Filter Bubble effect may be misguided.

In this paper, we focus on measuring Google Search, as it is the most popular search engine. However, our methodology is Web service agnostic, and could be repeated on other search engines like Bing or Google News Search.

At a high-level, our methodology is to execute carefully controlled queries on Google Search to identify what user features trigger personalization. Each experiment follows a similar pattern: first, create x Google accounts that each vary by one specific feature. Second, execute q identical queries from each account, once per day for d days. Save the results of each query. Finally, compare the results of the queries to determine whether the same results are being served in the same order to each account. If the results vary between accounts, then the changes can be attributed to personalization linked to the given experimental feature. Note that we run some experimental treatments *without* Google accounts (e.g., to simulate users without Google accounts).

Sources of Noise. Despite the simplicity of the high-level experimental design, there are several sources of noise that can cause identical queries to return different results.

- **Updates to the Search Index:** Web search services constantly update their search indices. This means that the results for a query may change over time.
- **Distributed Infrastructure:** Large-scale Web search services are spread across geographically diverse datacenters. Our tests have shown that different datacenters may return different results for the same queries. It is likely that these differences arise due to inconsistencies in the search index across datacenters.
- **Geolocation:** Search engines use the user's IP address to provide localized results [51]. Thus, searches from different subnets may receive different results.
- **A/B Testing:** Web search services sometimes conduct A/B testing [24], where certain results are altered to measure whether users click on them more often. Thus, there may be a certain level of noise independent of all other factors.

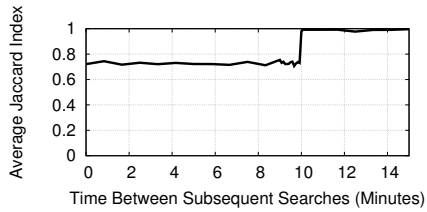


Figure 3: Overlap of results when searching for "test" followed by "touring" compared to just "touring" for different waiting periods.

The Carry-Over Effect. One particular source of noise comes from the influence of one search on subsequent searches. In other words, if a user searches for query A, and then searches for query B, the results for B may be influenced by the previous search for A. We term this phenomenon the *carry-over effect*. Prior research on user intent while searching has shown that sequential queries from a user are useful for refining search results [5, 40], so it is not surprising that Google Search leverages this feature.

An example of carry-over is shown in Figure 2. In this test, we search for "hawaii" and then immediately search for "urban outfitters" (a clothing retailer). We conducted the searches from a Boston IP address, so the results include links to the Urban Outfitters store in Boston. However, because the previous query was "hawaii," results pertaining to Urban Outfitters in Hawai'i are also shown.

To determine how close in time search queries must be to trigger carry-over, we conduct a simple experiment. We first pick different pairs of queries (e.g., "gay marriage" and "obama"). We then start two different browser instances: in one we search for the first query, wait, and then for the second query, while in the other we search only for the second query. We repeat this experiment with different wait times, and re-run the experiment 50 times with different query pairs. Finally, we compare the results returned in the two different browser instances for the second term.

The results of this experiment are shown in Figure 3 for the terms "test" and "touring" (other pairs of queries show similar results). The carry-over effect can be clearly observed: the results share, on average, seven common results (out of 10) when the interval between the searches is less than 10 minutes (in this case, results pertaining to Turing Tests are included). After 10 minutes, the carry-over effect disappears. Thus, in all experiments in the following sections, we wait at least 11 minutes between subsequent searches in order to avoid any carry-over effects. In our testing, we observed carry-over for both logged in users and users without Google accounts.

Controlling Against Noise. In order to mitigate measurements errors due to these factors, we perform a num-

ber of steps (some borrowed from [10]): *First*, all of our queries are executed by the normal Google Search webpage, rather than Google's Search API. It has been shown that search engine APIs sometimes return different results than the standard webpage [4]. *Second*, all of our machines execute searches for the same query at the same time (i.e., in lock-step). This eliminates differences is query results due to temporal effects. This also means that each of our Google accounts has exactly the same search history at the same time. *Third*, we use static DNS entries to direct all of our query traffic to a specific Google IP address. This eliminates errors arising from differences between datacenters. *Fourth*, we wait 11 minutes in-between subsequent queries to avoid carry-over. As shown in Figure 3, an 11 minute wait is sufficient to avoid the majority of instances of carry-over. *Fifth*, unless otherwise stated, we send all of the search queries for a given experiment from the same /24 subnet. Doing so ensures that any geolocation would affect all results equally.

Sixth, we include a *control account* in each of our experiments. The control account is configured in an identical manner to one other account in the given experiment (essentially, we run one of the experimental treatments twice). By comparing the results received by the control and its duplicate, we can determine the baseline level of noise in the experiment (e.g., noise caused by A/B testing). Intuitively, the control should receive exactly the same search results as its duplicate because they are configured identically, and perform the same actions at the same time. If there is divergence between their results, it must be due to noise.

3.2 Implementation

Our experiments are implemented using custom scripts for PhantomJS [28]. We chose PhantomJS because it is a full implementation of the WebKit browser, meaning that it executes JavaScript, manages cookies, *etc.* Thus, using PhantomJS is significantly more realistic than using custom code that does not execute JavaScript, and it is more scalable than automating a full Web browser (e.g., Selenium [42]).

On start, each PhantomJS instance logs in to Google using a separate Google account, and begins issuing queries to Google Search. The script downloads the first page of search results for each query. The script waits 11 minutes in-between searches for subsequent queries.

During execution, each PhantomJS instance remains persistent in memory and stores all received cookies. After executing all assigned queries, each PhantomJS instance closes and its cookies are cleared. The Google cookies are recreated during the next invocation of the experiment when the script logs in to its assigned Google account. All of our experiments are designed to complete in ≈24 hours.

All instances of PhantomJS are run on a single machine. We modified the `/etc/hosts` file of this machine so that Google DNS queries resolve to a specific Google IP address. We use SSH tunnels to forward traffic from each PhantomJS instance to a unique IP address in the same /24 subnet.

All of our experiments were conducted in fall of 2012. Although our results are representative for this time period, they may not hold in the future, since Google is constantly tweaking their personalization algorithms.

Google Accounts. Unless otherwise specified, each Google account we create has the same profile: 27 year old, female. The default User-Agent we use is Chrome 22 on Windows 7. As shown in Section 5.2, we do not observe any personalization of results based on these attributes.

We manually crafted each of our Google accounts to minimize the likelihood of Google automatically detecting them. Each account was given a unique name and profile image. We read all of the introductory emails in each account's Gmail inbox, and looked at any pending Google+ notifications. To the best of our knowledge, none of our accounts were banned or flagged by Google during our experiments.

Category	Examples	No.
Tech	Gadgets, Home Appliances	20
News	Politics, News Sources	20
Lifestyle	Apparel Brands, Travel Destinations, Home and Garden	30
Quirky	Weird Environmental, What-Is?	20
Humanities	Literature	10
Science	Health, Environment	20
Total		120

Table 1: Categories of search queries used in our experiments.

3.3 Search Queries

In our experiments, each Google account searches for a specific list of queries. It is fundamental to our research that we select a list of queries that has both breadth and impact. Breadth is vital, since we do not know which queries Google personalizes results for. However, given that we cannot test all possible queries, it is important that we select queries that real people are likely to use.

As shown in Table 1, we use 120 queries divided equally over 12 categories in our experiments. These queries were chosen from the 2011 Google Zeitgeist [14], and WebMD [48]. Google Zeitgeist is published annually by Google, and highlights the most popular search queries from the previous calendar year. We chose these queries for two reasons: first, they cover a broad range of categories (breadth). Second, these queries are popular by definition, i.e., they are guaranteed to impact a large number of people.

The queries from Google Zeitgeist cover many important areas. 10 queries are political (e.g., "Obama Jobs Plan", "2012 Republican Candidates") and 10 are related to news sources (e.g., "USA Today News"). Personalization of political and news-related searches are some of the most contentious issues raised in Eli Pariser's book on the Filter Bubble effects [26]. Furthermore, several categories are shopping related (e.g., gadgets, apparel brands, travel destination). As demonstrated by Orbitz, shopping related searches are prime targets for personalization [21].

One critical area that is not covered by Google Zeitgeist is health-related queries. To fill this gap, we chose ten random queries from WebMD's list of popular health topics [48].

4. REAL-WORLD PERSONALIZATION

We begin by measuring the extent of personalization that users are seeing today. Doing so requires obtaining access to the search results observed by real users; we therefore conducted a simple user study.

4.1 Collecting Real-World Data

We posted a task on Amazon's Mechanical Turk (AMT), explaining our study and offering each user $2.00 to participate. Participants were required to 1) be in the United

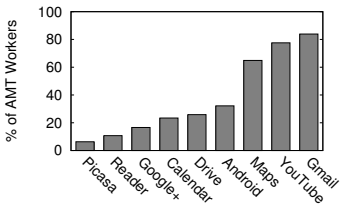


Figure 4: Usage of Google services by AMT workers.

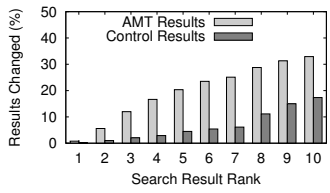


Figure 5: % of AMT and control results changed at each rank.

Most Personalized	Least Personalized
gap	what is gout
hollister	dance with dragons
hgtv	what is lupus
boomerang	gila monster facts
home depot	what is gluten
greece	ipad 2
pottery barn	cheri daniels
human rights	psoriatic arthritis
h2o	keurig coffee maker
nike	maytag refrigerator

Table 2: Top 10 most/least personalized queries.

States, 2) have a Google account, and 3) be logged in to Google during the study.[1] Users who accepted the task were instructed to configure their Web browser to use a HTTP proxy controlled by us. Then, the users were directed to visit a Web page that automatically performed 80 Google searches. 50 of the queries were randomly chosen from the categories in Table 1, while 30 were chosen by us.

The HTTP proxy serves several functions. *First*, the proxy records Google Search's HTML responses to the users' queries. *Second*, each time the proxy observes a user making a query, it executes two PhantomJS scripts. Each script logs in to Google and executes the same exact query as the user. The results served to the scripts act as the control, allowing us to compare results from a real user (who Google has collected extensive data on) to fresh accounts (that have minimal Google history). *Third*, the proxy controls for noise in two ways: 1) by executing user queries and the corresponding scripted queries in parallel, and 2) forwarding all Google Search traffic to a hard-coded Google IP address.

Although the proxy is necessary to control for noise, there is a caveat to this technique. Queries from AMT users must be sent to `http://google.com`, whereas the controls use `https://google.com`. The reason for this issue is that HTTPS Google Search rejects requests from proxies, since they could indicate a man-in-the-middle attack. Unfortunately, result pages from HTTP Google Search include a disclaimer explaining that some types of search personalization are disabled for HTTP results. Thus, our results from AMT users should be viewed as a lower bound on possible personalization.

AMT Worker Demographics. In total, we recruited 200 AMT workers, each of whom answered a brief demographic survey. Our participants self-reported to residing in 43 different U.S. states, and range in age from 12 to >48 (with a bias towards younger users). Figure 4 shows the usage of Google services by our participants: 84% are Gmail users, followed by 76% that use Google Maps. These survey results demonstrate that our participants 1) come from a broad sample of the U.S. population, and 2) use a wide variety of Google services.

4.2 Results

We now pose the question: *how often do real users receive personalized search results?* To answer this question, we compare the results received by AMT users and the corresponding control accounts. Figure 5 shows the percentage of results that differ at each rank (i.e., result 1, result 2, *etc.*) when we compare the AMT results to the control results, and the control results to each other. Intuitively, the percent change between the controls is the noise floor; any change above the noise floor when comparing AMT results to the control can be attributed to personalization.

There are two takeaways from Figure 5. First, we observe extensive personalization of search results. On average, across all ranks, AMT results showed an 11.7% *higher* likelihood of differing from the control result than the controls results did from each other. This additional difference can be attributed to personalization. Second, top ranks tend to be less personalized than bottom ranks.

To better understand how personalization varies across queries, we list the top 10 most and least personalized queries in Table 2. The level of personalization per query is calculated as the probability of AMT results equaling the control results, minus the probability of the control results equaling each other. Large values for this quantity indicate large divergence between AMT and control results, as well as low noise (i.e., low control/control divergence).

As shown in Table 2, the most personalized queries tend to be related to companies and politics (e.g., "greece", "human rights" or "home depot"). Digging into the individual results, we observe a great deal of personalization based on location. Even though all of the AMT users' requests went through our proxy and thus appeared to Google as being from the same IP address, Google Search returned results that are specific to other locations. This was especially common for company names, where AMT users received different store locations. In contrast, the least personalized results in Table 2 tend to be factual and health related queries.

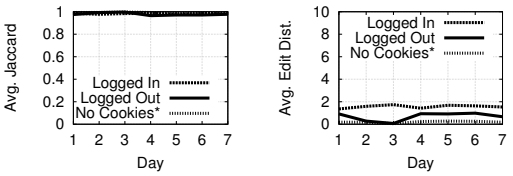
5. PERSONALIZATION FEATURES

In the previous section, we observed personalization for real users on Google Search. We now examine which user features Google Search uses to personalize results. Although we cannot possibly enumerate and test all possible features, we can investigate likely candidates. Table 3 lists the different demographic profiles that our experiments emulate during experiments.

5.1 Measuring Personalization

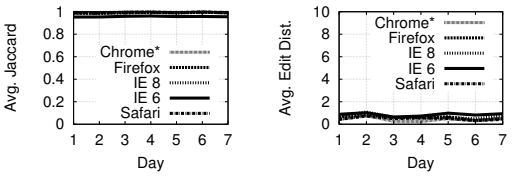
When comparing the list of search results for test and control accounts, we use two metrics to measure personalization. First, we use Jaccard Index, which views the result

[1]This study was conducted under Northeastern University IRB protocol #12-08-42; all personally identifiable information was removed from the dataset.



(a) Jaccard Index (b) Edit Distance

Figure 6: Results for the cookie tracking experiments.



(a) Jaccard Index (b) Edit Distance

Figure 7: Results for the browser experiments.

lists as sets and is defined as the size of the intersection over the size of the union. A Jaccard Index of 0 represents no overlap between the lists, while 1 indicates they contain the same results (although not necessarily in the same order).

To measure reordering, we use edit distance. To calculate edit distance, we compute the number of list elements that must be inserted, deleted, substituted, or swapped (i.e., the Damerau-Levenshtein distance [7]) to make the test list identical to the control list. For example, if the test account receives the result list [`a.com`, `b.com`, `c.com`] and the control receives the list [`c.com`, `b.com`] for the same query, then the edit distance is 2 (one insertion and one swap).

5.2 Basic Features

We begin our experiments by focusing on features associated with a user's browser, their physical location, and their Google profile. For each experiment, we create $x + 1$ fresh Google accounts, where x equals the number of possible values of the feature we are testing in that experiment, plus one additional control account. For example, in the Gender experiment, we create 4 accounts: one "male," one "female," one "other," and one additional "female" as a control. We execute $x+1$ instances of our PhantomJS script for each experiment, and forward the traffic to $x + 1$ unique endpoints via SSH tunnels. Each account searches for all 120 of our queries, and we repeat this process daily for seven days.

Basic Cookie Tracking. In this experiment, the goal is to compare the search results for users who are logged in to a Google account, not logged in to Google, and who do not support cookies at all. Google is able to track the logged in and logged out users, since Google Search places tracking cookies on all users, even if they do not have a Google account. The user who does not support cookies receives a new tracking cookie after every request to Google, and we confirm that the identifiers in these cookies are unique on every request. However, it is unknown whether Google is able to link these new identifiers together behind-the-scenes (e.g., by using the user's IP address as a unique identifier).

Category	Feature	Tested Values
Tracking	Cookies	Logged In, Logged Out, No Cookies
User-Agent	OS	Win. XP, Win. 7, OS X, Linux
	Browser	Chrome 22, Firefox 15, IE 6, IE 8, Safari 5
Geo-location	IP Address	MA, PA, IL, WA, CA, UT, NC, NY, OR, GA
Google Account	Gender	Male, Female, Other
	Age	15, 25, 35, 45, 55, 65
Search History, Click History, and Browsing History	Gender	Male, Female
	Age	<18, 18-24, 25-34, 35-44, 45-54, 55-64, ≥65
	Income	$0-50K, $50-100K, $100-150K, >$150K
	Education	No College, College, Grad School
	Ethnicity	Caucasian, African American, Asian, Hispanic

Table 3: User features evaluated for effects on search personalization.

To conduct this experiment, we use four instances of PhantomJS. The first two completely clear their cookies after every request. The third account logs in to Google and persists cookies normally. The fourth account does not log in to Google, and also persists cookies normally.

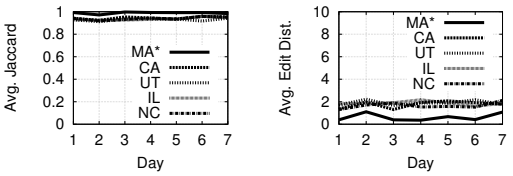
Figure 6(a) shows the average Jaccard Index for each account type (logged in/logged out/no cookies) across all search queries when compared to the control (no cookies). In all of our figures, we place a * on the legend entry that corresponds to the control test, i.e., two accounts that have identical features. We see from Figure 6(a) that the results received by users are not dependent on whether they support cookies, or their login state with Google. However, just because the results are the same, does not mean that they are returned in the same order.

To examine how the order of results changes, we plot the average edit distance between each account type versus the control in Figure 6(b). We observe that a user's login state and cookies do impact the order of results from Google Search. The greatest difference is between users who are logged in versus users that clear their cookies. Logged in users receive results that are reordered in two places (on average) as compared to users with no cookies. Logged out users also receive reordered results compared to the no cookie user, but the difference is smaller. The results in Figure 6 give the first glimpse of how Google alters search results for different types of users.

Browser User-Agent. Next, we examine whether the user's choice of browser or Operating System (OS) can impact search results. To test this, we created 11 Google accounts and assigned each one a different "User-Agent" string. As shown in Table 3, we encoded user-agents for 5 browsers and 4 OSs. Chrome 22 and Windows 7 serve as the controls.

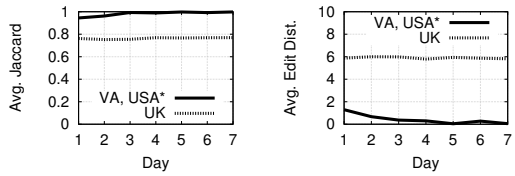
Figure 7 shows the results for our browser experiments. Unlike the cookie tracking experiment, there is no clear differentiation between the different browsers and the control experiment. The results for different OSs are similar, and we omit them for brevity. Thus, we do not observe search personalization based on user-agent strings.

IP address Geolocation. Next, we investigate whether Google Search personalizes results based on users' physical location. To examine this, we create 11 Google accounts and run our test suite while forwarding the traffic



(a) Jaccard Index (b) Edit Distance

Figure 8: Results for the geolocation experiments.



(a) Jaccard Index (b) Edit Distance

Figure 9: Results for inferred location experiments.

through SSH tunnels to 10 geographically diverse PlanetLab machines. These PlanetLab machines are located in the US states shown in Table 3. Two accounts forward through the Massachusetts PlanetLab machine, as it is the control.

Figure 8 shows the results of our location tests. There is a clear difference between the control and all the other locations. The average Jaccard Index for non-control tests is 0.91, meaning that queries from different locations generally differ by one result. The difference between locations is even more pronounced when we consider result order: the average edit distance for non-control accounts is 2.12.

These results reveal that Google Search does personalize results based on the user's geolocation. One example of this personalization can be seen by comparing the MA and CA results for the query "pier one" (a home furnishing store). The CA results include a link to a local news story covering a store grand opening in the area. In contrast, the MA results include a Google Maps link and a CitySearch link that highlight stores in the metropolitan area.

Inferred Geolocation. During our experiments, we observed one set of anomalous results from experiments that tunneled through Amazon EC2. In particular, 9 machines out of 22 rented from Amazon's North Virginia datacenter were receiving heavily personalized results, versus the other 13 machines, which showed no personalization. Manual investigation revealed that Google Search was returning results with `.co.uk` links to the 9 machines, while the 13 other machines received zero `.co.uk` links. The 9 machines receiving UK results were all located in the same /16 subnet.

Figure 9 shows some of the results for this anomaly. Although we could not determine why Google Search believes the 9 machines are in the UK (we believe it is due to an incorrect IP address geolocation database), we did confirm that this effect is independent of the Google account. As a result, we did not use EC2 machines as SSH tunnel endpoints for any of the results in this paper.

Google Account Attributes. In our next pair of tests, we examine whether Google Search uses demographic information from users' Google accounts to personalize results. Users must provide their gender and age when they sign up for a Google account, which means that Google Search could leverage this information to personalize results.

To test this hypothesis, we created Google accounts with specific demographic qualities. As shown in Table 3, we created "female," "male," and "other" accounts (these are the 3 choices Google gives during account sign-up), as well as accounts with ages 15 to 65, in increments of 10 years. The control account in the gender tests is female, while the control in the age tests is 15.

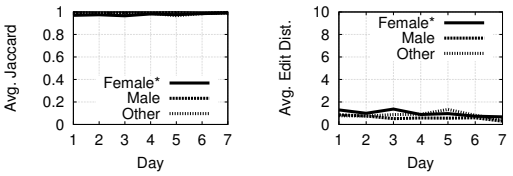
The results for the gender test are presented in Figure 10. We do not observe personalization based on gender in our experiments. Similarly, we do not observe personalization based on profile age, and we omit the results for brevity.

5.3 Historical Features

We now examine whether Google Search uses an account's history of activity to personalize results. We consider three types of historical actions: prior searches, prior searches where the user clicks a result, and Web browsing history.

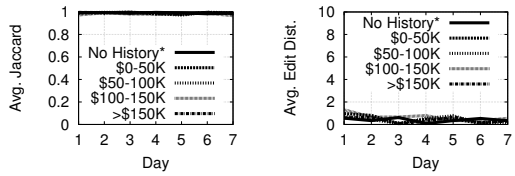
To create a plausible series of actions for different accounts, we use data from Quantcast, a Web analytics and advertising firm. Quantcast publishes a list of top websites (similar to Alexa) that includes the *demographics* of visitors to sites [30], broken down into the 20 categories shown in Table 3. Quantcast assigns each website a score for each demographic, where scores >100 indicate that the given demographic visits that website more frequently than average for the Web. The larger the score, the more heavily weighted the site's visitors are towards a particular demographic.

We use the Quantcast data to drive our historical experiments. In essence, our goal is to have different accounts "act" like a member of each of Quantcast's demographic groups. Thus, for each of our three experiments, we create 22 Google



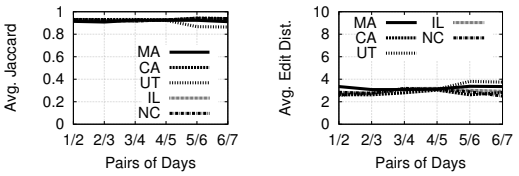
(a) Jaccard Index (b) Edit Distance

Figure 10: Results for the Google Profile: Gender experiments.



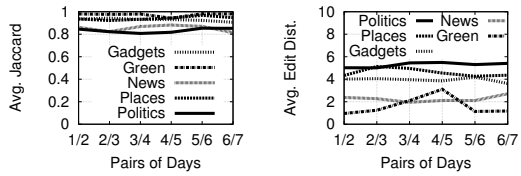
(a) Jaccard Index (b) Edit Distance

Figure 11: Results for the search history: income level experiments.



(a) Jaccard Index (b) Edit Distance

Figure 12: Day-to-day consistency of results for the geolocation experiments.



(a) Jaccard Index (b) Edit Distance

Figure 13: Day-to-day consistency within search query categories for the geolocation test.

accounts, two of which only run the 120 control queries, and 20 of which perform actions (i.e., searching, searching and clicking, or Web browsing) based on their assigned demographic before running the 120 control queries. For example, one account builds Web browsing history by visiting sites that are frequented by individuals earning >$150k per year. Each account is assigned a different Quantcast demographic, and chooses new action targets each day using weighted random selection, where the weights are based on Quantcast scores. For example, the >$150k browsing history account chooses new sites to browse each day from the corresponding list of URLs from Quantcast.

Search History. First, we examine whether Google Search personalizes results based on search history. Each day, the 20 test accounts search for 100 demographic queries before executing the standard 120 queries. The query strings are constructed by taking domains from the Quantcast top-2000 that have scores >100 for a particular demographic and removing subdomains and top level domains (e.g., `www.amazon.com` becomes "amazon").

Figure 11 shows the results of the search history test for four income demographics. The "No History" account does not search for demographic queries, and serves as the control. All accounts receive approximately the same search results, thus we do not observe personalization based on search history. This observation holds for all of the demographic categories we tested, and we omit the results for brevity.

Search-Result-Click History. Next, we examine whether Google Search personalizes results based on the search results that a user has clicked on. We use the same methodology as for the search history experiment, with the addition that accounts click on the search results that match their demographic queries. For example, an account that searches for "amazon" would click on the result for `amazon.com`. Accounts will go through multiple pages of search results to find the correct link for a given query.

The results of the click history experiments are the same as for the search history experiments. There is little difference between the controls and the test accounts, regardless of demographic. Thus, we do not observe personalization based on click history, and we omit the results for brevity.

Browsing History. Finally, we investigate whether Google Search personalizes results based on Web browsing history (i.e., by tracking users on third-party Web sites). In these experiments, each account logs into Google and then browses 5 random pages from 50 demographically skewed websites each day. We filter out websites that do not set Google cookies (or Google affiliates like DoubleClick), since Google cannot track visits to these sites. Out of 1,587 unique domains in the Quantcast data that have scores >100, 700 include Google tracking cookies.

The results of the browsing history experiments are the same as for search history and click history: regardless of demographic, we do not observe personalization. We omit these results for brevity.

Discussion. We were surprised that the history-driven tests did not reveal personalization on Google Search. One explanation for this finding is that account history may only impact search results for a brief time window, i.e., carry-over is the extent of history-driven personalization on Google Search. As future work, we plan on conducting longer lasting history-driven experiments to confirm our findings.

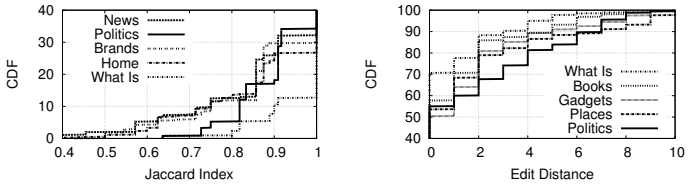
6. QUANTIFYING PERSONALIZATION

In the previous section we demonstrate that Google Search personalization occurs based on 1) whether the user is logged in and 2) the location of the searching machine. In this section, we dive deeper into the data from our synthetic experiments to better understand how personalization impacts search results. First, we examine the temporal dynamics of search results. Next, we investigate the amount of personalization in different categories of queries. Finally, we examine the rank of personalized search results to understand whether certain positions are more volatile than others.

6.1 Temporal Dynamics

In this section, we examine the temporal dynamics of results from Google Search to understand how much results from Google Search change day-to-day, and whether personalized results are more or less volatile than non-personalized search results. To measure the dynamics of Google Search over time, we compute the Jaccard Index and edit distance for search results from subsequent days. Figure 12 shows the day-to-day dynamics for our geolocation experiment. The x-axis shows which two days of search results are being compared, and each line corresponds to a particular test account.

Figure 12 reveals three facts about Google Search. First, the lines in Figures 12 are roughly horizontal, indicating that the rate of change in the search index is constant. Second, we observe that there is more reordering over time than new results: average Jaccard Index is 0.9, while average edit distance is 3. Third, we observe that both of these trends are consistent across all of our experiments, irrespective of whether the results are personalized. This indicates that personalization does not increase the day-to-day volatility of search results.



(a) Jaccard Index (b) Edit Distance

Figure 14: Differences in search results for five query categories.

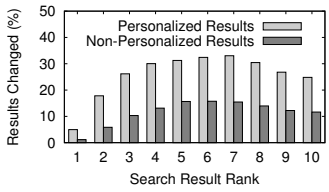


Figure 15: The percentage of results changed at each rank.

Dynamics of Query Categories. We now examine the temporal dynamics of results across different categories of queries. As shown in Table 1, we use 12 categories of queries in our experiments. Our goal is to understand whether each category is equally volatile over time, or whether certain categories evolve more than others.

To understand the dynamics of query categories, we again calculate the Jaccard Index and edit distance between search results from subsequent days. However, instead of grouping by experiment, we now group by query category. Figure 13 shows the day-to-day dynamics for query categories during our geolocation experiment. Although we have 12 categories in total, Figure 13 only shows the 1 least volatile, and 4 most volatile categories, for clarity. The results for all other experiments are similar to the results for the geolocation test, and we omit them for brevity.

Figure 13 reveals that the search results for different query categories change at different rates day-to-day. Figure 13(a) shows that there are more new results per day for "politics" and "news" queries. Similarly, Figure 13(b) shows that queries for "politics," "news," and "places" all exhibit above average reordering each day. This reflects how quickly information in these categories changes on the Web. In the case of "places," the reordering is due to location specific news items that fluctuate daily. In contrast, search queries for factual categories like "what is" and "green" (environmentally friendly topics) are less volatile over time.

6.2 Personalization of Query Categories

We now examine the relationship between different categories of search queries and personalization. In Section 5, we demonstrate that Google Search does personalize search results. However, it remains unclear whether all categories of queries receive equal amounts of personalization.

To answer this question, we plot the cumulative distribute of Jaccard Index and edit distance for each category in Figure 14. These results are calculated over all of our experiments (i.e., User-Agent, Google Profile, geolocation, *etc.*) for a single day of search results. For clarity, we only include lines for the 1 most stable category (i.e., Jaccard close to 1, edit distance close to 0), and the 4 least stable categories.

Figure 14 demonstrates that Google Search personalizes results for some query categories more than others. For example, 82% of results for "what is" queries are identical, while only 43% of results for "gadgets" are identical. Overall, "politics" is the most personalized query category, followed by "places" and "gadgets." CDFs calculated over other days of search results demonstrate nearly identical results.

6.3 Personalization and Result Ranking

In this section, we focus on the volatility of results from Google Search at each rank, with rank 1 being the first result on the page and rank 10 being the last result. Understanding the impact of personalization on top ranked search results is critical, since eye-tracking studies have demonstrated that users rarely scroll down to results "below the fold" [3, 12, 13, 20]. Thus, we have two goals: 1) to understand whether certain ranks are more volatile in general, and 2) to examine whether personalized search results are more volatile than non-personalized results.

To answer these questions, we plot Figure 15, which shows the percentage of results that change at each rank. To calculate these values, we perform a pairwise comparison between the result at rank $r \in [1, 10]$ received by a test account and the corresponding control. We perform comparisons across all tests in all experiments, across all seven days of measurement. This produces a total number of results that are changed at each rank r, which we divide by the total number of results at rank r to produce a percentage. The personalized results come from the logged in/logged out and geolocation experiments; all others are non-personalized.

Figure 15 reveals two interesting features. First, the results on personalized pages are significantly more volatile than the results on non-personalized pages. The result changes on non-personalized pages represent the noise floor of the experiment; at every rank, there are more than twice as many changes on personalized pages. Second, Figure 15 shows that the volatility at each rank is not uniform. Rank 1 exhibits the least volatility, and the volatility increases until it peaks at 33% in rank 7. This indicates that Google Search is more conservative about altering results at top ranks.

Given the extreme importance placed on rank 1 in Google Search, we now delve deeper into the 5% of cases where the result at rank 1 changes during personalized searches. In

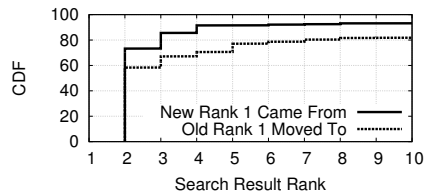


Figure 16: Movement of results to and from rank 1 for personalized searches.

each instance where the rank 1 result changes, we compare the results for the test account and the control to determine 1) what was the *original rank* of the result that moved to rank 1, and 2) what is the *new rank* of the result that used to be at rank 1. Figure 16 plots the results of this test. In the vast majority of cases, the rank 1 and 2 results switch places: 73% of new rank 1 results originate from rank 2, and 58% of old rank 1 results move to rank 2. Overall, 93% of new rank 1 results come from the first page of results, while 82% of old rank 1 results remain somewhere on the first result page. However, neither CDF sums to 100%, i.e., there are cases where the new rank 1 result does not appear in the control results and/or the old rank 1 result disappears completely from the test results. The latter case is more common, with 18% of rank 1 results getting evicted completely from the first page of results.

7. RELATED WORK

Comparing Search Engines. Several studies have examined the differences between results from different search engines. Two studies have performed user studies to compare search engines [2, 44]. Although both studies uncover significant differences between competing search engines, neither study examines the impact of personalization. Sun et al. propose a method for visualizing different results from search engines that is based on expected weighted Hoeffding distance [37]. Although this technique is very promising, it does not scale to the size of our experiments.

Personalization. Personalized search has been extensively studied in the literature [8, 19, 23, 27, 29, 36, 39, 43]. Dou et al. provide a comprehensive overview of techniques for personalizing search [6]. They evaluate many strategies for personalizing search, and conclude that mining user click histories leads to the most accurate results. In contrast, user profiles have low utility. The authors also note that personalization is not useful for all types of queries.

Other features besides click history have been used to power personalized search. Three studies leverage geographic location to personalize search [41,50,51]. Two studies have shown that user demographics can be reliably inferred from browsing histories, which can be useful for personalizing content [11,16]. To our knowledge, only one study has investigated privacy-preserving personalized search [49]. Given growing concerns about the Filter Bubble effects, this area seems promising for future research.

Several studies have looked at personalization on systems other than search. Two studies have examined personalization of targeted ads on the Web [10,47]. One study examines discriminatory pricing on e-commerce sites, which is essentially personalization of prices [22].

8. CONCLUDING DISCUSSION

Over the past few years, we have witnessed a trend of personalization in numerous Internet-based services, including Web search. While personalization provides obvious benefits for users, it also opens up the possibility that certain information may be unintentionally hidden from users. Despite the variety of speculation on this topic, to date, there has been little quantification of the basis and extent of personalization in Web search services today.

In this paper, we take the first steps towards addressing this situation by introducing a methodology for measuring personalization on Web search engines. Our methodology controls for numerous sources of noise, allowing us to accurately measure the extent of personalization. We applied our methodology to real Google accounts recruited from AMT and observe that 11.7% of search results show differences due to personalization. Using artificially created accounts, we observe that measurable personalization is caused by 1) being logged in to Google and 2) making requests from different geographic areas.

However, much work remains to be done: we view our results as a first step in providing transparency for users of Web search and other Web-based services. In the paragraphs below, we discuss a few of the issues brought up by our work, as well as promising directions for future research.

Scope. In this paper, we focus on queries to US version of the Google Web Search. All queries are in English, and are drawn from topics that are primarily of interest to US residents. We leave the examination of Google sites in other countries and other languages to future work.

Incompleteness. As a result of our methodology, we are only able to identify positive instances of personalization; we cannot claim the absence of personalization, as we may not have considered other dimensions along which personalization could occur. However, the dimensions that we chose to examine in this paper are the most obvious ones for personalization (considering how much prior work has looked at demographic, location-based, and history-based personalization). Given that any form of personalization is a moving target, we aim to continue this work by running our data collection for a longer time, looking at additional categories of Web searches, examining searches from mobile devices, and looking at other user behaviors (e.g., using services like Gmail, Google+, and Google Maps). We also plan on examining the impact of mechanisms that may disable personalization (e.g., opting-out of personalization on Google Search, and enabling Do-Not-Track headers).

Generality. The methodology that we develop is not specific to Google Web Search. The sources of noise that we control for are present in other search engines (e.g., Bing, Google News Search) as well as other Web-based services (e.g., Twitter search, Yelp recommendations, *etc.*). We plan on applying our methodology to these and other search services to quantify personalization of different types.

Impact. In this paper, we focused on quantifying literal differences in search results, e.g., `a.com` is different from `b.com`. However, we do not address the issue of semantic differences, i.e., do `a.com` and `b.com` contain different information content? If so, what is the impact of these differences? While semantic differences and impact are challenging to quantify, we plan to explore natural language processing and user studies as a first step.

Open Source. We make all of the crawling and parsing code, as well as the Google Web Search data from Section 5, available to the research community at

`http://personalization.ccs.neu.edu/`

Acknowledgements

We thank the anonymous reviewers for their helpful comments. This research was supported by NSF grants IIS-0964465 and CNS-1054233, and an Amazon Web Services in Education Grant.

9. REFERENCES

[1] Personalized Search Graduates from Google Labs. *News From Google Blog*, 2005. http://bit.ly/Tndpgf.

[2] J. Bar-Ilan, K. Keenoy, E. Yaari, and M. Levene. User rankings of search engine results. *J. Am. Soc. Inf. Sci. Technol.*, 58(9), 2007.

[3] E. Cutrell and Z. Guan. What are you looking for?: an eye-tracking study of information usage in web search. *CHI*, 2007.

[4] F. Crown and M. L. Nelson. Agreeing to Disagree: Search Engines and Their Public Interfaces. *JCDL*, 2007.

[5] Z. Cheng, B. Gao, and T.-Y. Liu. Actively Predicting Diverse Search Intent from User Browsing Behaviors. *WWW*, 2010.

[6] Z. Dou, R. Song, and J.-R. Wen. A Large-scale Evaluation and Analysis of Personalized Search Strategies. *WWW*, 2007.

[7] F. J. Damerau. A technique for computer detection and correction of spelling errors. *CACM*, 7(3), 1964.

[8] J. T. S. T. Dumais and E. Horvitz. Personalizing search via automated analysis of interests and activities. *SIGIR*, 2005.

[9] H. Green. Breaking Out of Your Internet Filter Bubble. *Forbes*, 2011. http://onforb.es/oYWbDf.

[10] S. Guha, B. Cheng, and P. Francis. Challenges in Measuring Online Advertising Systems. *IMC*, 2010.

[11] S. Goel, J. M. Hofman, and M. I. Sirer. Who Does What on the Web: A Large-scale Study of Browsing Behavior. *ICWSM*, 2012.

[12] Z. Guan and E. Cutrell. An eye tracking study of the effect of target rank on web search. *CHI*, 2007.

[13] L. A. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in WWW search. *SIGIR*, 2004.

[14] Google Zeitgeist. http://www.googlezeitgeist.com.

[15] B. Horling and M. Kulick. Personalized Search for Everyone. *Google Official Blog*, 2009. http://bit.ly/71RcmJ.

[16] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen. Demographic Prediction Based on User's Browsing Behavior. *WWW*, 2007.

[17] M. Hines. Google Takes Searching Personally. *CNet*, 2004. http://cnet.co/V37pZD.

[18] How Rick Santorum's 'Google Problem' Has Endured. http://n.pr/wefdnc.

[19] F. Liu, C. Yu, and W. Meng. Personalized web search by mapping user queries to categories. *CIKM*, 2002.

[20] L. Lorigo, B. Pan, H. Hembrooke, T. Joachims, L. Granka, and G. Gay. The influence of task and gender on search and evaluation behavior using Google. *Inf. Process. Manage.*, 42(4), 2006.

[21] D. Mattioli. On Orbitz, Mac Users Steered to Pricier Hotels. *Wall Street Journal*, 2012. http://on.wsj.com/LwTnPH.

[22] J. Mikians, L. Gyarmati, V. Erramilli, and N. Laoutaris. Detecting Price and Search Discrimination on the Internet. *HotNets*, 2012.

[23] A. Pretschner and S. Gauch. Ontology based personalized search. *ICTAI*, 1999.

[24] A. Pansari and M. Mayer. This is a test. This is only a test. *Google Official Blog*, 2006. http://bit.ly/Ldbb0.

[25] B. Pan, H. Hembrooke, T. Joachims, L. Lorigo, G. Gay, and L. Granka. In Google We Trust: Users' Decisions on Rank, Position, and Relevance. *J. Comp. Med. Comm.*, 12(3), 2007.

[26] E. Pariser. *The Filter Bubble: What the Internet is Hiding from You*. Penguin Press, 2011.

[27] J. Pitkow, H. Schütze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, and T. Breuel. Personalized search. *CACM*, 45(9), 2002.

[28] PhantomJS. http://phantomjs.org.

[29] F. Qiu and J. Cho. Automatic Identification of User Interest for Personalized Search. *WWW*, 2006.

[30] Quantcast. Top Sites for the United States. 2012. http://www.quantcast.com/top-sites.

[31] F. Roesner, T. Kohno, and D. Wetherall. Detecting and Defending Against Third-Party Tracking on the Web. *NSDI*, 2012.

[32] A. Singhal. Some Thoughts on Personalization. *Google Inside Search Blog*, 2011. http://bit.ly/tJS4xT.

[33] A. Singhal. Search, Plus Your World. *Google Official Blog*, 2012. http://bit.ly/yUJnCl.

[34] D. Sullivan. Bing Results Get Local and Personalized. *Search Engine Land*, 2011. http://selnd.com/hY4djp.

[35] D. Sullivan. Why Google "Personalizes" Results Based on Obama Searches But Not Romney. *Search Engine Land*, 2012. http://selnd.com/PyfvvY.

[36] J.-T. Sun, H.-J. Zeng, H. Liu, Y. Lu, and Z. Chen. CubeSVD: A Novel Approach to Personalized Web Search. *WWW*, 2005.

[37] M. Sun, G. Lebanon, and K. Collins-Thompson. Visualizing Differences in Web Search Algorithms using the Expected Weighted Hoeffding Distance. *WWW*, 2010.

[38] N. Singer. The Trouble with the Echo Chamber Online. *The New York Times*, 2011. http://nyti.ms/jcTih2.

[39] X. Shen, B. Tan, and C. Zhai. Implicit user modeling for personalized search. *CIKM*, 2005.

[40] Y. Shen, J. Yan, S. Yan, L. Ji, N. Liu, and Z. Chen. Sparse Hidden-Dynamics Conditional Random Fields for User Intent Understanding. *WWW*, 2011.

[41] L. A. M. J. Silva. Relevance Ranking for Geographic IR. *GIR*, 2006.

[42] Selenium. http://selenium.org.

[43] B. Tan, X. Shen, and C. Zhai. Mining long-term search history to improve search accuracy. *KDD*, 2006.

[44] L. Vaughan. New measurements for search engine evaluation proposed and tested. *Inf. Process. Manage.*, 40(4), 2004.

[45] A. Witten. Google's New Privacy Policy. *Google Official Blog*, 2012. http://bit.ly/wVr4mF.

[46] M. Wines. Google to Alert Users to Chinese Censorship. *The New York Times*, 2012. http://nyti.ms/JRhGZS.

[47] C. E. Wills and C. Tatar. Understanding What They Do with What They Know. *WPES*, 2012.

[48] WebMD 2011 Year in Health. http://on.webmd.com/eBPFxH.

[49] Y. Xu, B. Zhang, Z. Chen, and K. Wang. Privacy-Enhancing Personalized Web Search. *WWW*, 2007.

[50] B. Yu and G. Cai. A query-aware document ranking method for geographic information retrieval. *GIR*, 2007.

[51] X. Yi, H. Raghavan, and C. Leggetter. Discovering Users' Specific Geo Intention in Web Search. *WWW*, 2009.

[52] comScore August 2012 U.S. Search Results. http://bit.ly/ThGnOc.

Bibliography

[And13] Android Developers. Connectivity — Bluetooth. http://developer.android.com/guide/topics/connectivity/bluetooth.html, 2013. [Online; accessed 03-March-2013].

[API+11] Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland. The Social fMRI: Measuring, Understanding, and Designing Social Mechanisms in the Real World. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 445–454. ACM, 2011.

[App13] Apple Support. Bluetooth glossary of terms. http://support.apple.com/kb/HT3894#discoverable, 2013. [Online; accessed 03-March-2013].

[AS03] Daniel Ashbrook and Thad Starner. Using gps to learn significant locations and predict movement across multiple users. *Personal Ubiquitous Comput.*, 7(5):275–286, October 2003.

[BAB+10] Nilanjan Banerjee, Sharad Agarwal, Paramvir Bahl, Ranveer Chandra, Alec Wolman, and Mark Corner. Virtual Compass: Relative Positioning to Sense Mobile Social Interactions. In *Proceedings of the 8th international conference on Pervasive Computing*, Pervasive'10, pages 1–21, Berlin, Heidelberg, 2010. Springer-Verlag.

[BCG+09] Duygu Balcan, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J. Ramasco, and Alessandro Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases.

Proceedings of the National Academy of Sciences, 106(51):21484–21489, December 2009.

[Ben07] A. Bensky. *Wireless positioning technologies and applications*. Artech House, Inc., 2007.

[BHG06] D Brockmann, L Hufnagel, and T Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, January 2006.

[BHWS10] D.M. Bullock, R. Haseman, J.S. Wasson, and R. Spitler. Automated measurement of wait times at airport security. *Journal of the Transportation Research Board*, 2177(-1):60–68, 2010.

[BKS79] H. R. Bernard, Peter D. Killworth, and Lee Sailer. Informant accuracy in social network data IV: a comparison of clique-level structure in behavioral and cognitive network data. *Social Networks*, 2(3):191–218, January 1979.

[Bla13] BlackBerry Enterprise Solution security. Security Technical Overview — BlackBerry Devices with Bluetooth Technology. http://docs.blackberry.com/en/admin/deliverables/1854/Security_for_BlackBerry_Devices_with_Bluetooth_Wireless_Technology.pdf, 2013. [Online; accessed 03-March-2013].

[BN12] Brian Ball and M. E. J. Newman. Friendship networks and social status. *CoRR*, abs/1205.6822, 2012.

[BZ] Michael Barbaro and Tom Zeller. A Face Is Exposed for AOL Searcher No. 4417749. *The New York Times*.

[CBB+10] Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. Dynamics of Person-To-Person Interactions from Distributed RFID Sensor Networks. 1159(arXiv:1007.3680):6, Jul 2010. Comments: see also http://www.sociopatterns.org.

[CBC+10a] David J. Crandall, Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher, and Jon Kleinberg. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 2010.

[CBC+10b] David J. Crandall, Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher, and Jon Kleinberg. Inferring social ties from geographic coincidences. *Proc. of the National Academy of Sciences*, 107(52):22436–22441, 2010.

[Cho04] Tanzeem Khalid Choudhury. *Sensing and Modeling Human Networks*. PhD thesis, 2004. AAI0806108.

[CMSO12] Iacopo Carreras, A. Matic, Piret Saar, and V. Osmani. Comm2Sense: Detecting Proximity Through Smartphones. *PerMoby Workshop, part of IEEE PerCom '12 Conference*, March 2012.

[CTH+10] Justin Cranshaw, Eran Toch, Jason Hong, Aniket Kittur, and Norman Sadeh. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, Ubicomp '10, pages 119–128, New York, NY, USA, 2010. ACM.

[Dah05] David B. Dahl. Sequentially-allocated merge-split sampler for conjugate and nonconjugate Dirichlet process mixture models. Technical report, Texas A&M University, 2005.

[DDLM12] Manlio De Domenico, Antonio Lima, and Mirco Musolesi. Interdependence and predictability of human mobility and social interactions. In *Proceedings of the Nokia Mobile Data Challenge Workshop*, Newcastle, United Kingdom, June 2012.

[Dun92] R.I.M. Dunbar. Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22(6):469 – 493, 1992.

[EP06a] N. Eagle and A. Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006.

[EP06b] Nathan Eagle and Alex Pentland. Eigenbehaviors: Identifying Structure in Routine. Technical report, IN PROC. OF UBICOMP Ó6, 2006.

[EPL09] Nathan Eagle, Alex (Sandy) Pentland, and David Lazer. Inferring Friendship Network Structure by Using Mobile Phone Data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009.

[GBB+02] William G. Griswold, Robert Boyer, Steven W. Brown, Tan Minh Truong, Ezekiel Bhasker, Gregory R. Jay, and R. Benjamin Shapiro. Using Mobile Technology to Create Opportunitistic Interactions on a University Campus. In *UBICOMP 2002 ON SUPPORTING SPONTANEOUS INTERACTION IN UBIQUITOUS COMPUTING SETTINGS, TECHNICAL REPORT CS20020724, COMPUTER SCIENCE AND ENGINEERING*, 2002.

[GHB08] M.C. Gonzalez, C.A. Hidalgo, and A.L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[Goo12] Google and Ipsos OTX Media CT. Our Mobile Planet: Denmark. http://www.thinkwithgoogle.com/insights/library/studies/our-mobile-planet-Denmark, 2012. [Online; accessed 03-March-2013].

[HAA+09] J. Hansen, A. Alapetite, H. Andersen, L. Malmborg, and J. Thommesen. Location-based services and privacy in airports. *Human-Computer Interaction–INTERACT 2009*, pages 168–181, 2009.

[HCS+05] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and C. Diot. Pocket switched networks and human mobility in conference environments. In *Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking*, pages 244–251. ACM, 2005.

[HH09] S. Hay and R. Harle. Bluetooth tracking without discoverability. *Location and Context Awareness*, pages 120–137, 2009.

[HMH11] T.J. Hansen, M. Morup, and L.K. Hansen. Non-parametric co-clustering of large scale sparse bipartite networks on the gpu. In *IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1 –6, 2011.

[HWB10] R.J. Haseman, J.S. Wasson, and DM Bullock. Real time measurement of work zone travel time delay and evaluation metrics using bluetooth probe tracking. *Journal of the Transportation Research Board*, 2010.

[IDC13] IDC. IDC — Press Release. http://www.idc.com/getdoc.jsp?containerId=prUS23946013#.UTNTGaF35t1, 2013. [Online; accessed 03-March-2013].

[IEE] IEEE. Public OUI and Company ID Assignments. http://standards.ieee.org/develop/regauth/oui/.

[JLB05] Ravi Jain, Dan Lelescu, and Mahadevan Balakrishnan. Model T: an Empirical Model for User Registration Patterns in a Campus Wireless LAN. In *Proceedings of the 11th annual international conference on Mobile computing and networking*, MobiCom '05, pages 170–184, New York, NY, USA, 2005. ACM.

[JLJ+10] B.S. Jensen, J.E. Larsen, K. Jensen, J. Larsen, and L.K. Hansen. Estimating human predictability from mobile sensor data. In *IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP)*, pages 196–201, 2010.

[JN04] Sonia Jain and Radford M. Neal. A split-merge markov chain monte carlo procedure for the dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13(1):158–182, 2004.

[KBD+10] N. Kiukkonen, J. Blom, O. Dousse, D. Gatica-Perez, and J. Laurila. Towards rich mobile phone datasets: Lausanne data collection campaign. *Proc. ICPS*, 2010.

[Kel10] D. Kelly. *Minimal Infrastructure Radio Frequency Home Localisation Systems*. PhD thesis, National University of Ireland, 2010.

[KH04] John Krumm and Ken Hinckley. The NearMe Wireless Proximity Server. In *In Proceedings of Ubicomp: Ubiquitous Computing*, pages 283–300. Springer, 2004.

[KN12] Mikkel Baun Kjærgaard and Petteri Nurmi. Challenges for Social Sensing Using WiFi Signals. In *Proceedings of the 1st ACM workshop on Mobile systems for computational social science*, MCSS '12, pages 17–21, New York, NY, USA, 2012. ACM.

[KOP+10] Vassilis Kostakos, Eamonn O'Neill, Alan Penn, George Roussos, and Dikaios Papadongonas. Brief encounters: Sensing, modeling and visualizing urban mobility and copresence networks. *ACM Trans. Comput.-Hum. Interact.*, 17(1):2:1–2:38, April 2010.

[Kra] David Krackhardt. Constraints on the interactive organization as an ideal type. In *The Post-Bureaucratic Organization*.

[KTG+06] Charles Kemp, Joshua B. Tenenbaum, Thomas L. Griffiths, Takeshi Yamada, and Naonori Ueda. Learning systems of concepts with an infinite relational model. In *Proc. of the National AAAI Conf. on Artificial Intelligence*, 2006.

[KWRT12] Mikkel Baun Kjærgaard, Martin Wirz, Daniel Roggen, and Gerhard Tröster. Mobile Sensing of Pedestrian Flocks in Indoor Environments Using WiFi Signals. In *Proceedings of the IEEE International Conference on Pervasive Computing and Communications*, 2012.

[LL07] Ninghui Li and Tiancheng Li. t-closeness: Privacy beyond k-anonymity and ℓ-diversity. In *In Proc. of IEEE 23rd Int'l Conf. on Data Engineering (ICDE'07*, 2007.

[LM95] G.Y. Liu and Jr. Maguire, G.Q. A predictive mobility management algorithm for wireless mobile computing and communications. In *Universal Personal Communications. 1995. Record., 1995 Fourth IEEE International Conference on*, pages 268–272, 1995.

[LPA+09] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Social science: Computational social science. *Science*, 323(5915):721–723, February 2009.

[LS11] Shu Liu and Aaron Striegel. Accurate Extraction of Face-To-Face Proximity Using Smartphones and Bluetooth. In *Computer Communications and Networks (ICCCN), 2011 Proceedings of 20th International Conference on*, pages 1–5. IEEE, 2011.

[MGP10] R. Montoliu and D. Gatica-Perez. Discovering human places of interest from multimodal mobile phone data. In *Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia*, page 12. ACM, 2010.

[MKGV07] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. ℓ-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 1(1), March 2007.

[MSZ04] S. Maslov, K. Sneppen, and A. Zaliznyak. Detection of topological patterns in complex networks: correlation profile of the internet. *Physica A: Statistical Mechanics and its Applications*, 333:529–540, 2004.

[NFC13] NFC Forum. About NFC — Key Benefits of NFC. http://www.nfc-forum.org/aboutnfc/, 2013. [Online; accessed 25-February-2013].

[NS08] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, SP '08, pages 111–125, Washington, DC, USA, 2008. IEEE Computer Society.

[OKK+06] Eamonn O'Neill, Vassilis Kostakos, Tim Kindberg, Ava Fatah gen. Schiek, Alan Penn, Danaë Stanton Fraser, and Tim Jones. Instrumenting the city: developing methods for observing and understanding the digital cityscape. In *Proceedings of the 8th*

international conference on Ubiquitous Computing, UbiComp'06, pages 315–332, Berlin, Heidelberg, 2006. Springer-Verlag.

[OWK+09] Daniel Olguín Olguín, Benjamin N. Waber, Taemie Kim, Akshay Mohan, Koji Ara, and Alex Pentland. Sensible Organizations: Technology and Methodology for Automatically Measuring Organizational Behavior. *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS-PART B: CYBERNETICS*, pages 43–55, 2009.

[PBK06] B. S. Peterson, R. O. Baldwin, and J. P. Kharoufeh. Bluetooth inquiry time characterization and selection. *IEEE Transactions on Mobile Computing*, 5(9):1173–1187, 2006.

[Pit06] J. Pitman. *Combinatorial stochastic processes*, volume 1875. Springer-Verlag, 2006.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[Ros02] Barbara Rosenstein. Video use in social science research and program evaluation. *International Journal of Qulitative Methods*, 2002.

[RZZB12] Gyan Ranjan, Hui Zang, Zhi-Li Zhang, and Jean Bolot. Are Call Detail Records Biased for Sampling Human Mobility? *SIGMOBILE Mob. Comput. Commun. Rev.*, 16(3):33–44, December 2012.

[SLH+11] H. Stange, T. Liebig, D. Hecker, G. Andrienko, and N. Andrienko. Analytical workflow of monitoring human mobility in big event settings using bluetooth. In *Proc. of the 3rd ACM SIGSPATIAL Int.l Workshop on Indoor Spatial Awareness*, pages 51–58. ACM, 2011.

[SLL+13] A. Stopczynski, J.E Larsen, S. Lehmann, Dynowski L., and Fuentes M. Participatory Bluetooth Sensing: A Method for Acquiring Spatio-Temporal Data about Participant Mobility and Interactions at Large Scale Events. In *Pervasive Computing and Communications Workshops, 2013. PerCom Workshops '13*, 2013.

[SNM11] Salvatore Scellato, Anastasios Noulas, and Cecilia Mascolo. Exploiting place features in link prediction on location-based social

networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 1046–1054, New York, NY, USA, 2011. ACM.

[SQBB10] C. Song, Z. Qu, N. Blumm, and A.L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018, 2010.

[Swe01] Latanya Sweeney. Computational Disclosure Control – a primer on data privacy protection. Technical report, Massachusetts Institute of Technology, 2001.

[Swe02] Latanya Sweeney. *k*-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, October 2002.

[TSK06] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2006.

[VDNVdW10] M. Versichele, M. Delafontaine, T. Neutens, and N. Van de Weghe. Potential and implications of bluetooth proximity-based tracking in moving object research. In *1st Int. workshop on movement pattern analysis (MPA) in conj. with the 6th Int. conf. on Geographic Information Science*, 2010.

[VNDVdW12] M. Versichele, T. Neutens, M. Delafontaine, and N. Van de Weghe. The use of bluetooth for analysing spatiotemporal dynamics of human movement at mass events: A case study of the ghent festivities. *Applied Geography*, 32(2):208–220, 2012.

[Win13] Windows Phone. Pair my phone with a Bluetooth accessory. http://www.windowsphone.com/en-US/How-to/wp8/start/pair-my-phone-with-a-bluetooth-accessory, 2013. [Online; accessed 03-March-2013].

[WPP+12] Martin Whitehead, Tom Phillips, Mark Page, Maria Molina, and Charlotte Wood. European mobile industry observatory 2011. http://www.gsma.com/publicpolicy/wp-content/uploads/2012/04/emofullwebfinal.pdf, 2012. [Online; accessed 07-March-2013].

[WPS+11] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-Laszlo Barabasi. Human Mobility, Social Ties, and Link Prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 1100–1108, New York, NY, USA, 2011. ACM.

[XTYK06] Zhao Xu, Volker Tresp, Kai Yu, and Hans-Peter Kriegel. Learning infinite hidden relational models. *Uncertainty in Artificial Intelligence (UAI2006)*, 2006.