Privacy-respecting Features in Large Collections of Personal Data

M.Sc. Thesis

Benjamin Holm Glaas
s123862

Technical University of Denmark
Department of Applied Mathematics
and Computer Science

Supervisors:
Sune Lehmann Jørgensen
Enys Mones

June 16, 2017

Abstract

Privacy has become an important aspect of today's technology and can be difficult to control as vast amounts of personal data are being generated every day. Recent research has shown that in 95 % of the investigated cases using only 4 spatio-temporal points are enough to uniquely identify individuals when combining the points with a database of otherwise anonymized mobility data. While mobility data evidently reveals a lot about a person, other types of personal data has not yet been investigated. Furthermore, up until now individuals have been treated equally when identifying them, while indeed they are different. This thesis investigates how users differ in large databases containing mobility data as well as social data, using findability measures that characterizes each individual. Additionally, labelled data is investigated, yielding a lower resolution of data than what has previously been discussed in the literature to find out how it affects users' findability. An in-depth analysis of missing data is also being performed, where missing data is treated in different scenarios. The results indicate that using multiple labels makes identification of users easier and using combined spatial and social labels in turn makes identification easier than using plain spatial labels. Furthermore, an analysis shows that certain time bins of either the hour of the day or the week are more important than others when identifying users. These analyses give a deeper understanding of how to protect one's personal privacy in large databases.

Resumé

Privatlivet er blevet et vigtigt aspekt i nutidens teknologi og kan være svært at kontrollere da enorme mængder data bliver genereret hver dag. Ny forskning har vist at i 95 % af de undersøgte tilfælde er 4 spatielle-temporale punkter nok til at unikt identificere individer når disse punkter kombineres med en database, som ellers indeholder anonymiseret mobilitets data. Mens mobilitets data klart afslører meget om en person, er andre typer af personlig data endnu ikke blevet undersøgt. Endvidere er individer indtil nu kun blevet behandlet ens når de identificeres, mens de bestemt er forskellige. Dette speciale undersøger hvordan brugere adskiller sig fra hinanden i store databaser som indeholder mobilitets samt social data ved at kvantificere hvor let det er at finde den enkelte bruger. Desuden er labelled data undersøgt, som har en lavere opløsning af data end hvad der tidligere er blevet brugt i litteraturen for at finde ud af hvordan det påvirker brugeres evne til at blive fundet. En dybdegående analyse af manglende data bliver også udført, hvor manglende data bliver behandlet i forskellige scenarier. Resultaterne indikerer at hvis der bruges adskillige labels er brugere lettere at finde og brug af kombinerede rumlige og sociale labels gør det til gengæld lettere at finde brugere i forhold til almindelige rumlige labels. Endvidere viser en analyse at visse timer på dagen eller ugen er vigtigere end andre når det gælder om at identificere brugere. Disse analyser giver indsigt i hvordan ens personlige privatliv kan bevares i store databaser.

Table of Contents

Preface

This thesis was conducted at the Department of Applied Mathematics and Computer Science at the Technical University of Denmark in the period December 2016 to June 2017. The thesis concludes the Master of Science degree in Digital Media Engineering in fulfilment of the requirements for obtaining the degree. The scope of the thesis was outlined in collaboration with my supervisors Sune Lehmann and Enys Mones.

I would like to thank Sune Lehmann for supervising the project and helping with understanding the theory and results in depth, as well as helping with the textual contents of the thesis. Also, thank you to Enys Mones for co-supervising the project and helping with generating and understanding the datasets. Thank you to Ulf Aslak, who helped with pointers in the last rounds of writing. I thank my friends and family for support and finally thank you Marie, for always being there in times of distress.

Benjamin Glaas

Benjamin Holm Glaas
June 16, 2017

Chapter 1

Introduction

Vast amounts of data, estimated to 2.5 exabytes ($2.5 \cdot 10^9$ gigabytes) a day [1], are being produced all the time, and each individual is part of this data rich world. We are all contributing our share of data through the use of the internet, whether using our smart phone or pc. Data such as social interactions (Facebook statuses, Twitter tweets, etc.), GPS data and fitness tracking are all part of the bigger picture, but how hard is it really to find a specific individual in this vast amount of data? And how important is it to encrypt or anonymize the data if one does not want to be found?

It is clear that every individual wants to be private in some sense, but the amount of data being produced and stored *in the cloud* can coincide with this privacy. For example, some apps are collecting background data through the microphone and camera and some agencies collect massive amounts of information of individuals to sell their products accordingly to each user's preferences [2]. It has been discussed whether simply removing names and personal info in large datasets is enough to make sure that one cannot be identified. Individuals are for example very regular in their mobility, but still very unique in contrast to others, so just by looking at their whereabouts can lead to their identification. Tele-communication companies also regularly let their massive datasets be open to the public for either research or commercial purposes, containing only aggregated or masked data believing that this will protect their customers. However, mobile trajectories can still be easily restored even with this masking because of highly regular but still unique patterns of users [3].

At the moment, users cannot fully control their personal data they themselves generate. Generated metadata, that is *data about data*, from applications or services that keep preferences or logs of users cannot for the most part be accessed by the user and the user cannot control who may be given access to their metadata. These metadata records are also scattered between all services that the user has given consent to, yielding obsolete storage of records that may be equivalent for different services. A proposed solution is the openPDS (open Personal Data Store) framework, which stores all metadata about a user from different applications. In this framework, users can themselves control and access their metadata, as well as control who gets access to the information that is contained in the data store. This is useful for the services as well, as they don't need to collect and store the data themselves since it probably already exists in the data store. Then, when services need the data to find preferences for the user, they cannot access the raw data, but get aggregated data through a question-and-answer system called SafeAnswers [4]. Another idea of personal data storage which is built on the blockchain technology has also been proposed, where users also are in more control of their personal data. Blockchain is primarily used in connection with the bitcoin currency, which has been shown to be a secure way of storing data [5]. These proposals are excellent ideas to preserve personal data, however they have not yet been commercialized.

As individuals we all want to preserve our privacy, whether it be personal information we regard as sensitive or maybe privacy regarding our personal intimate space. In the aftermath of World War II, the United Nations General Assembly wrote their Universal Declaration of Human Rights, in which Article 12 states that [6]

> “*No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks.*”

From the article given by the United Nations, it calls to mind that the right to keep one’s privacy if wanted is a universal right on its own. The common term *Big Brother*, originating from the 1949 novel *Nineteen Eighty-Four* or *1984* by George Orwell - in which a totalitarian regime exists where everyone is being watched in their home, only being able to do what the government allows - is a popular example of a total breach of privacy, which of course is fictional. However, the book may be a criticism on how the author viewed the society at the time and the future development of such. Indeed the book has inspired many modern thinkings, and some people even draw comparisons to the reality in the book and modern day, believing we live in a similar society where the government has too much control [7]. Additionally, a 2015 American survey performed by Pew Research Center shows that 52 % of the test participants said they were "very concerned" or "concerned" about the governments’ surveillance of Americans’ data and electronic communications [8].

The definition of privacy has varied and developed over time, especially given the information era in which we currently live. A concrete definition of privacy is mostly a subjective matter, but even so, Himma and Tavani define different types of privacy [9, Ch. 6]:

1. Physical/accessibility privacy
2. Decisional privacy
3. Psychological/mental privacy
4. Informational privacy

Physical privacy is the definition of having privacy in the physical space. **Decisional privacy** can be described as the personal freedom to make plans, choices and decisions without having external interference. **Psychological privacy** is a kind of privacy that lets people have their intimate thoughts protected or not being manipulated by others. Finally, **informational privacy** is described as a privacy that concerns personal data stored in and communicated between electronic databases, as well as personal information exchanged between parties using e-mail, telephony and wireless communication devices. These kinds of informations can be about a persons daily activities, personal lifestyle, finances, medical history and academic achievement [9]. The last kind of privacy is the one that will be discussed in this thesis, analysing how one can keep their informational privacy in a world full of databases containing their personal information.

Informational privacy is very relevant for many people in the world, and it can easily be breached when one simply uses the World Wide Web. When browsing the Web, users set a so-called *fingerprint* with their browser, which is fairly easy to collect. Fingerprints hold information on features such as which browser is used, which default language is used, which time zone is used and which plug-ins are used. Certain web pages can show a user based on their browser fingerprints whether they are easy to identify or not [10]. A study shows that from the data they had collected when users visited a certain website, 83.6 % of these fingerprints were unique. The study also shows that iPhone and Android web browsers are not as unique as desktop browsers, mostly because there are not as many plug-ins available [11]. Using the fingerprints of users can breach their privacy on-line, where attackers can gain knowledge about users’ preferences and habits. Usually people want to be unique in their own way, but in the case of uniqueness in web browsers users are easy targets.

In other contexts, being unique can also be dangerous regarding informational privacy breaches. A study shows that around 99 % of the test subjects involved had a unique list of apps installed on their smart phone, and that knowing any four apps installed can uniquely identify a user 95 % of the time. This means, that if someone is identified, the attacker knows the rest of the apps installed on a user's smart phone. This leads to a privacy breach, since installed apps say a lot about users' behaviour in general and if services such as Twitter, which keeps track of all installed apps, would share this information to advertising companies even in an anonymized format, it would have consequences for the users [12]. Therefore, individuals who use modern technologies, which is close to everyone in technologically advanced countries, are easy victims of cybercrime without even knowing it.

In the sense of uniqueness, one does not even have to actively browse the web or install applications to be in risk of being identified. If one has a Google account and is logged in on their smart phone while having the location reporting setting turned on, Google will track the user's whereabouts [13]. This is not a privacy breach as such, since one can easily turn off the settings, but it has been shown that one can easily be identified by knowing only a few points in time where the person has been. It has been shown by de Montjoye et al. that four spatio-temporal points is enough to uniquely identify 95 % of the test subjects [14]. This means that if a user has been identified by only four spatio-temporal points, an attacker knows the whereabouts of the user for the rest of the observations in the dataset, including their home as this can easily be calculated as the location which most time is spent during the night. If the database holds years of records then the privacy breach is indeed disturbing as mobility records are very sensitive regarding users' behaviour.

In another study, de Montjoye et al. has shown that four spatio-temporal points are enough to uniquely identify 90 % of users in a dataset involving their credit card transactions, which indeed also carry sensitive information [15]. In both of the studies by de Montjoye et al., obvious identifiers such as names, age, gender, etc. have been removed, yielding it intuitively hard to find people in a very large dataset. However, one can easily get a hold of these 4 spatio-temporal points, either by seeing a post on social media such as Facebook, where someone for example has described a trip to the cinema or maybe a vacation, a geo-located tweet from Twitter or by eavesdropping on a conversation in public. Indeed, if an attacker gets a hold of points of users where they are significantly different than other users, the identifications become easy.

As discussed by de Montjoye et al., the resolution of the data impacts the privacy of users. For example, knowing the precise location of a user at a given time yields more information than only knowing whether they were at home or not. To justify this observation, this thesis takes inspiration from the work done by de Monjoye et al. and uses labelled data when identifying users, containing low-resolution data as opposed to the high-resolution data used in the papers. It has been shown before that the use of labels of the top N locations visited instead of raw GPS-values indeed creates more privacy-respecting features in the database [16], however previous work, including the work done by de Montjoye et al., only discuss results of aggregating over users. Since users are different with different behaviour in their mobilities, analyses concerning the *individual* qualities of users will be discussed in this thesis. This includes a measure of findability for each user, Non-negative Matrix Factorization (NMF) and likelihood methods to separate users into different groups as well as classification methods for classifying users based on their NMF features.

An analysis of which time bins are holding more important information than others when identifying users will also be given, which will be addressed as *temporal correlations*. For example, it may be that users are more identifiable when using information for certain hours of the day rather than others. Also, databases holding personal information obtained from smart phones or other services can hold values of missing data, either from not observing an event or from erroneous storage of records. Indeed, missing values can either have the meaning of an event not observed or the fact that no event was there, so this thesis will also discuss the importance of missing data. To do this, different scenarios when identifying users will be presented.

In this thesis a two labels mobility dataset containing labels of the top 1 location, which with high probability is the home of a user, and the location of *somewhere else*, will be analysed to compare to the results obtained by de Montjoye et al. Furthermore, a four labels mobility dataset will be discussed as well, holding the top 3 locations of users as well as the label *somewhere else* will be analysed to show the difference when increasing the resolution. Additionally, a two labels social dataset, where labels of whether a social event was present or not, is studied for comparison to the two labels mobility dataset. Finally, a four labels combined dataset will be used to compare to the four labels mobility dataset, where the combined labels are a combination of spatial and social labels.

For the analysis given in this thesis, numeric tools have been used. First of all, the Python programming language version 3.5 has been used with the Anaconda distribution [17]. The Python language makes it easy to import and handle large amounts of data with the *pandas* module as well as providing an easy and interpretable syntax. Within the Python language the IPython notebook has been used as an interface for writing and running the code and the *matplotlib* module has been used for plotting the results.

Chapter 2

Sensible DTU dataset

In this chapter the Sensible DTU dataset will be discussed. The study in which the data was obtained will be elaborated in Section 2.1 and each of the sub-datasets used in this thesis will be discussed as well. These sub-datasets include a mobility dataset discussed in Section 2.2 and a social dataset discussed in Section 2.3.

2.1 Sensible DTU study

The datasets used in this thesis are all sub-datasets obtained in the Sensible DTU study which took place at the Technical University of Denmark from 2013 to 2016 [18]. An initial deployment in 2012 was also made, but with fewer test participants [19].

The motivation behind the Sensible DTU study has primarily been to get multiple channel data, where earlier studies only has focused on single-channel data. To elaborate, earlier analysis on for example mobility data has only been conducted using primarily Call Detail Records (CDR) which are metadata records holding information on calls and text messages. CDR's hold cell tower locations, which can be used as a proxy for locations of users, however only within a 100 metres or so of the user. Using multiple channels, mobility records can be found from different sensors using a smart phone, for example WiFi hotspot scans, GPS locations and cell towers, combining different tools to get the most accurate and battery saving options available. Furthermore, combining different channels yield deeper understanding of how social gatherings are happening using Bluetooth scans, WiFi hotspot scans and Facebook events to find the social networks of users. Furthermore, questionnaires and qualitative data have been collected, where the qualitative data was obtained by an anthropologist observing social groups during and after the study hours.

In both deployments, the test subjects were undergraduates which started the project either a few weeks in the semester (2012) or in the beginning of the semester (2013). In 2012, around 200 smart phones were distributed between students in different study lines and in 2013 around 1000 smart phones were distributed. To ensure privacy in the study, the students had been informed on what their data would be used for. Furthermore, in the 2013 deployment, all participants had the opportunity to view their data with the same API as the researchers were using with a built-in data visualization tool. The participants were also able to keep in contact with the researchers through blogposts and social media [19].

Based on the data obtained in the study, a number of projects have been conducted. For example, one study has shown that it is easy to infer WiFi Access Points (APs) as a proxy for human mobility. Using only a few APs, one of the author's mobility could be described more than 90 % of the time. However, as WiFi scans are generally not seen as a threat to safety, being able to find the user's whereabouts solely based on APs can be a threat to one's privacy [20]. In another study, it has been shown that so-called communities, which can be described as dense groups of nodes in a network, can be found directly when

looking at different temporal "slices" [21]. Other studies based on the Sensible DTU study include work on epidemiology, privacy and other studies related to the so-called *computational social science* field [22]. For a full list of studies based on the Sensible DTU dataset, see [23].
In the following sections the datasets which are used in this thesis are described. The datasets include mobility labels and social labels.

2.2 Home/Work labels

In this section the **Home/Work** dataset will be discussed, showing general characteristics as well as briefly discussing missing data. The data is obtained by finding the location of a user by the best method available *at the moment*. This means that different methods can be used to find the location, either from GPS, WiFi hotspots or cell towers [19]. The dataset consists of mobility data for 835 users with hourly resolutions. The data is within two years, yielding 17520 time bins. It should also be mentioned that this dataset, in contrast to the other datasets used in this thesis, does not use specific time stamps, but simple indices. Therefore, the code used for the different datasets slightly differ, but not much. Instead of specific (longitude, latitude) coordinates, the data has been labelled 1 if the user is at their top most visited location, which can be deduced to the home of the user, and 0 if the user was not at home, where all obvious indicators such as name, gender, age, etc. have been removed. Missing data is indicated with a -1. This labelling will from now on be addressed as **Home/Work** labels.
In Figure 2.1 one can see an example of how much users are at home during the day. In Figure 2.1 one can see the behaviour of user 0 and user 1, as well as the mean of all users. It can be seen that some users are quite different when comparing user 0 and 1. The mean of all users seems plausible, namely that most users are mostly at home during the evening and night, which is not that different from user 0 and 1. In Figure 2.2 one can see how much users are at home during the hours of the week. From Figure 2.2 it is clear that there is a periodic behaviour of the users, which is quite intuitive. Users have their specific routine, and many users leave their home on Fridays to go and meet others. However, it is also clear that users are different, for example that user 0 is more at home on Fridays in general than user 1. It should be mentioned that this periodic behaviour is not clear for all users, as some users lack much

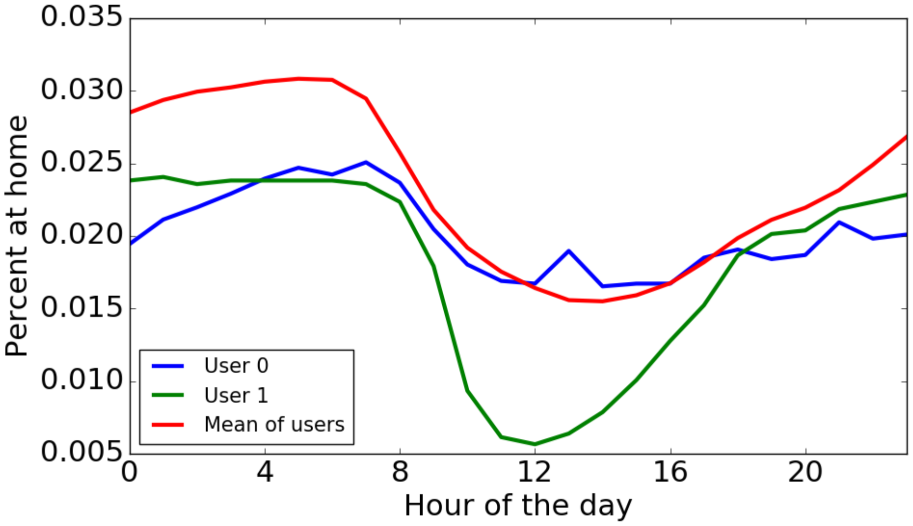

Figure 2.1: *Figure showing how many percent users have been home during the day. Here two different users as well as the mean of all users are shown. It can be seen that users can differ quite much when comparing user 0 and 1.*

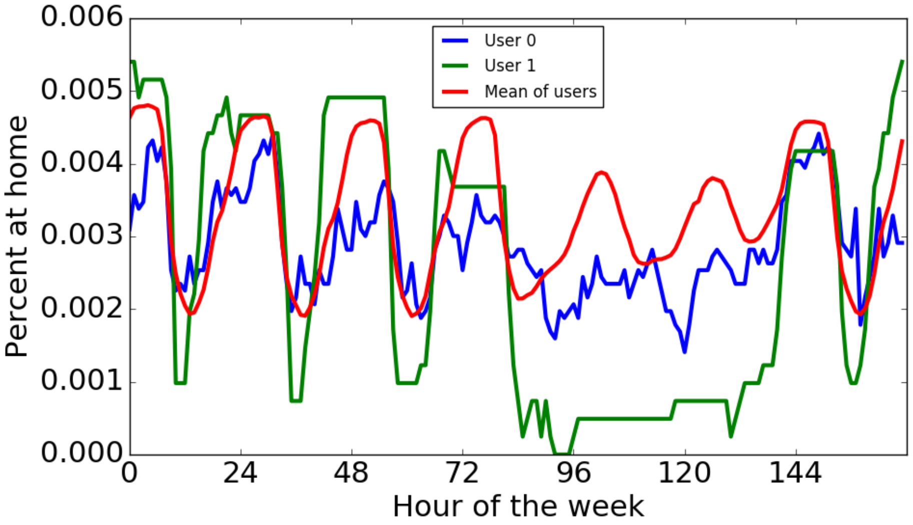

Figure 2.2: *Figure showing how many percent users have been home during the hours of the week. Here two different users as well as the mean of all users are shown. When comparing users 0 and 1, it can be seen that they are quite different, especially on Fridays and Saturdays.*

information, that is they have a lot of missing data. A brief discussion on missing data is presented in Section 2.2.1.

2.2.1 Missing data

The **Home/Work** dataset has a lot of missing data where each time bin and each user have missing data. In Figure 2.3 one can see a histogram showing the amount of missing data for all users. From Figure

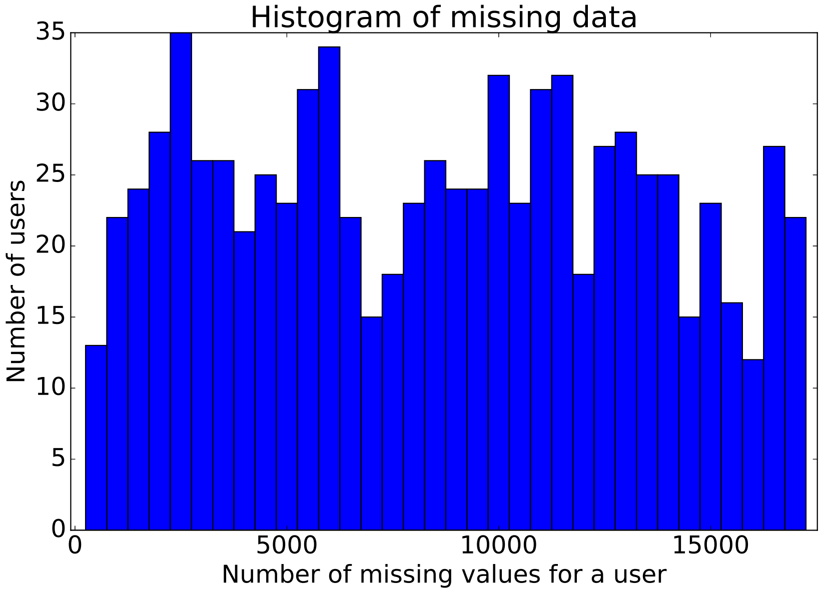

Figure 2.3: *Histogram showing the number of missing data for users. The mean number of missing data bins for all users is 8769.71, where the lowest for a user is 320 and the highest is 17519.*

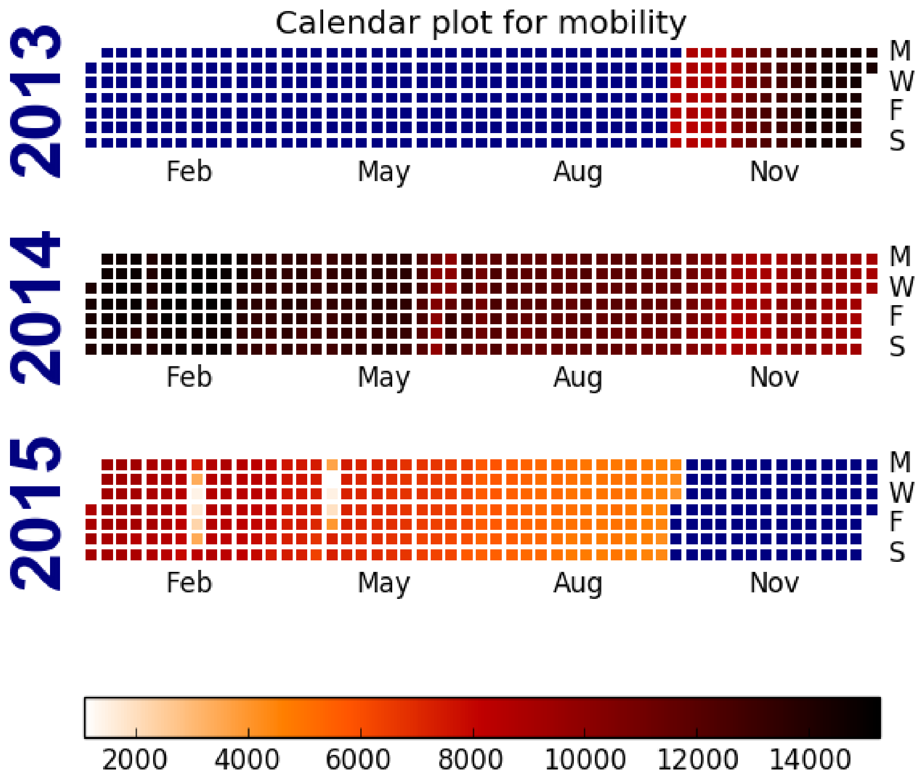

Figure 2.4: *Figure showing the number of available observations for each day of the study, where a blue color means no data. It is evident that the largest number of available observations are in the middle of the study where most test participants were present.*

2.3 one can see that the number of missing values for all users are quite different, with the lowest being 320 and the highest being 17519, meaning one user only has a single valid bin. The average number of missing values is found to be 8769.71. Additionally, Figure 2.4 shows the number of observations collected from each day during the time period of the study. In the figure, the blue color means *no data*. From Figure 2.4 we can see, that the largest number of available observations is in the middle of the study. This makes sense, since not all students have joined in the beginning of the study and not all users have been there till the end [19].
As missing data should be treated with caution and may be important when identifying users, different scenarios for handling missing data will be presented in Section 3.1. The analyses performed on the **Home/Work** dataset is seen in Chapter 4.

2.3 Social labels

The social datasets considered in this thesis consist of Facebook, call and sms observations. A 1 is given if there is a Facebook/call/sms event (observation) and a 0 if none, meaning there actually is no missing data even if an event is not recorded. However, a long sequence of 0's may occur because data could not be obtained in this time period, meaning that missing data still can occur, but it cannot easily be seen.

Table 2.1: *Number of users and observations for the three social datasets.*

	Facebook	Call	Sms
Number of users	864	1016	1016
Number of time stamps	20982	19830	29695

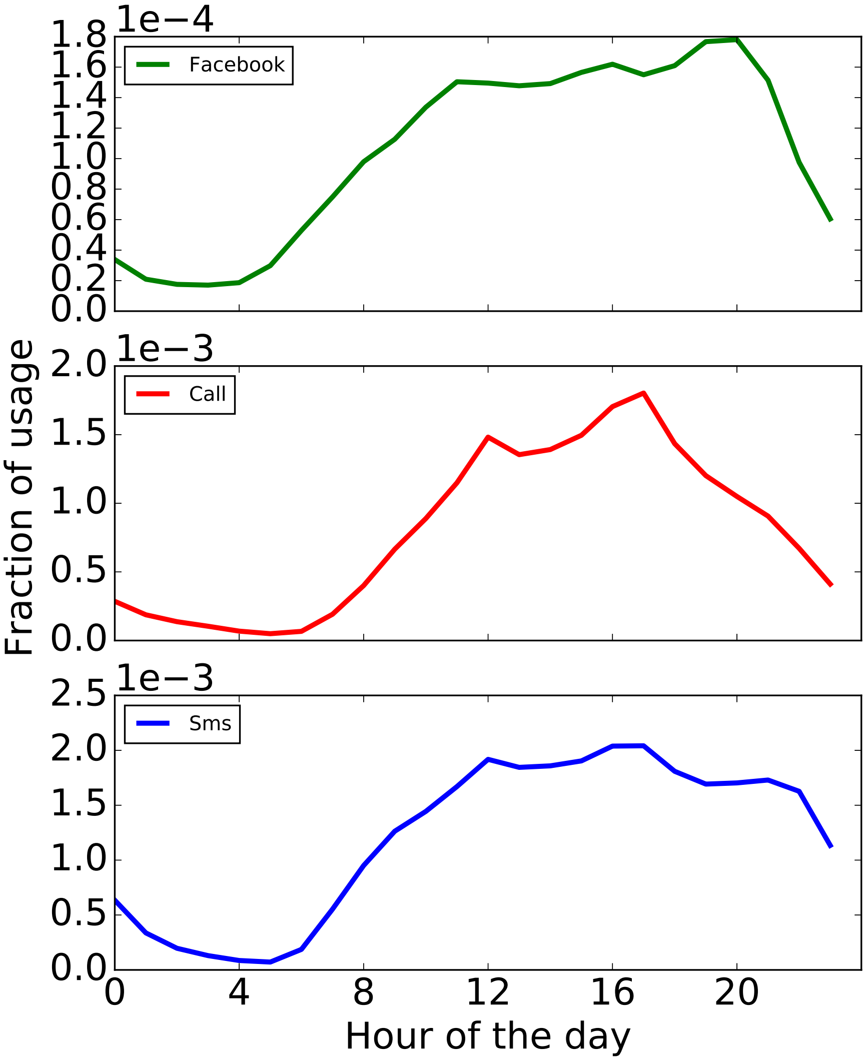

Figure 2.5: *Figure showing the percentage of usage for the three social datasets on a daily basis. It is noticeable that the fraction of Facebook observations is much lower (a factor of 10) than for the other two datasets. Furthermore, users are more socially active in the afternoon and evening in comparison to the morning and night time.*

The number of users and the number of time stamps for each dataset can be seen in Table 2.1. From Table 2.1 it can be seen that the number of users deviate, which may be caused by the different roll-outs of the study (see Section 2.1). In Figure 2.5, one can see the fraction of usage on a daily basis for the three social datasets. From Figure 2.5 it is clear that there are much fewer Facebook observations than

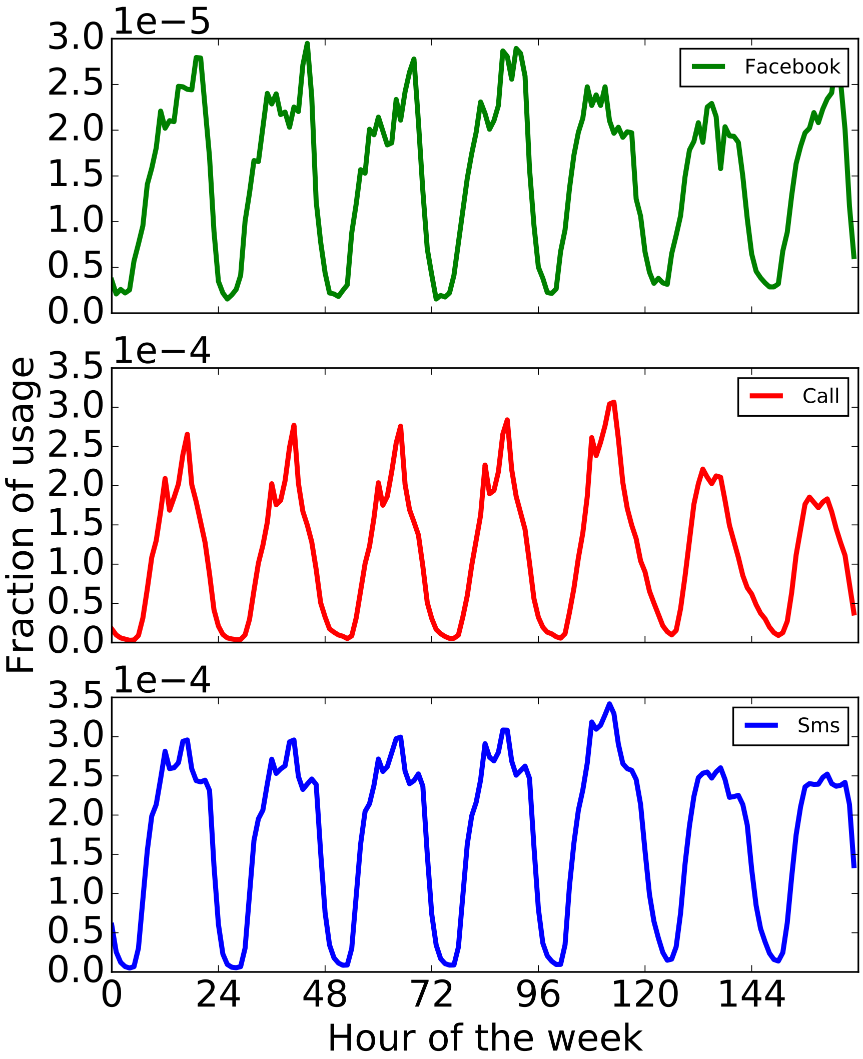

Figure 2.6: *Figure showing the percentage of usage for the three social datasets on a weekly basis. As for the daily basis, the fraction of Facebook observations is much lower (a factor of 10) than for the other two datasets and in general very low for all datasets.*

call and sms observations and in general very low fractions for all datasets on a daily basis. It is also clear that users are mostly socially active in the afternoon and evening and least active in the morning and night time. In Figure 2.6 one can see the fraction of usage on a weekly basis. From Figure 2.6 one can see that the fraction of Facebook observations is lower than for the other two datasets, which is the same case as in Figure 2.5. It can also be seen that there is a specific pattern, which resembles the daily pattern, that is users are more socially active in the afternoon and the least in the night time. From both Figure 2.5 and Figure 2.6 it is indicated that users have the same amount of calls as the amount of sms messages, which is kind of odd, since one would think that the amount of sms messages would be higher. However, since both call and sms services are much less used today, it may be valid that they are used equally, since internet services/apps are used for messaging instead of sms.

The three discussed datasets do not hold very much information though and early analysis shows that it is hard to identify people in each dataset separately. To get some interesting results and define a more free state of *being social*, the results from the analysis on the social labels given in Chapter 5 is based on a combination of the three datasets mentioned, which will be elaborated in Section 5.1. This dataset will be called the **dialogue** dataset.

2.4 Four spatial labels

The four labels spatial dataset considered in this section holds the labels of the top three locations of a user, where the fourth label is *another location*. This means that a label of 0 means that the user (typically) is at home and a label of 1 is the second most visited location, which may be work or in this case the test subjects being students, the university. A label of 2 means the third most visited location and a label of 3 is finally any other place. As for the **Home/Work** dataset, missing values are present and a placeholder label of -1 is used for this kind of data. The missing values of this dataset is (almost) the same as for the **Home/Work** dataset, seeing that only the labels should differ, so these will not be

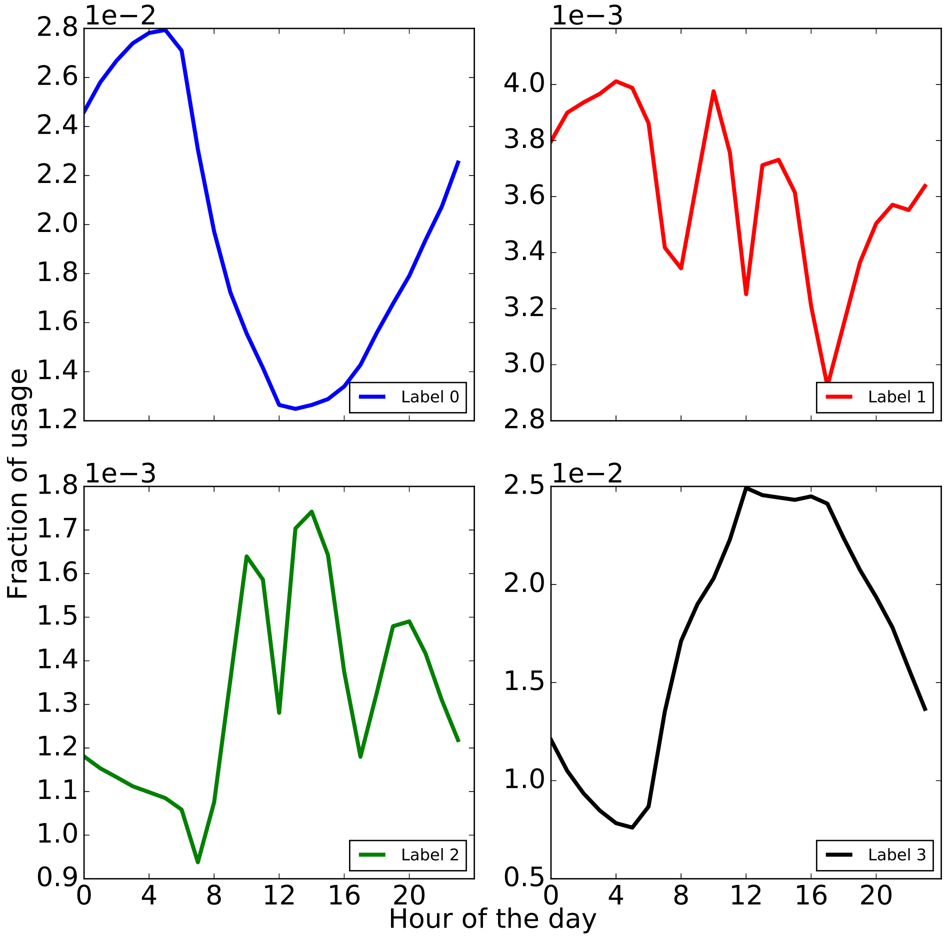

Figure 2.7: *Figure showing the percentage of usage for the labels 0 to 3 in the four labels mobility dataset. It can be seen that the most dominant locations of users are at home, which is label 0, and the location* ***other****, which is anywhere but the top three locations, which makes sense. Furthermore, the locations of labels 1 and 2 are a magnitude of 10 smaller than labels 0 and 3, and seems more unpredictable than labels 0 and 3. Here a daily resolution has been used.*

discussed. However, there are slight changes in the number of missing data in the two datasets, where the four labels dataset has some more missing values, but this is with high probability caused by the generation of the datasets. As this dataset was generated in the very end of the thesis, the methods of generating the datasets may have been altered, but the methods of finding the labels is out of the scope of this thesis. It is also suspected that the results are not changed drastically based on these alterations, so no further processing has been done. It should also be mentioned that a few of the users have two labels at the same timestamp, yielding the user practically two different places at once. This however is caused by the time frame, which is 1 hour, so the user can be at the two locations (or maybe more) within 1 hour. Therefore, if two labels are present, the lowest label number has been used.

The four labels spatial dataset has 835 users and 17520 time bins, yielding two years of data with hourly resolutions in correspondence with the **Home/Work** dataset. In Figure 2.7, one can see the fraction of usage for the 4 labels for a daily resolution. In the figure, only valid data bins have been used. From Figure 2.7 it can be seen that labels 0 and 3, that is being at home and somewhere else than the

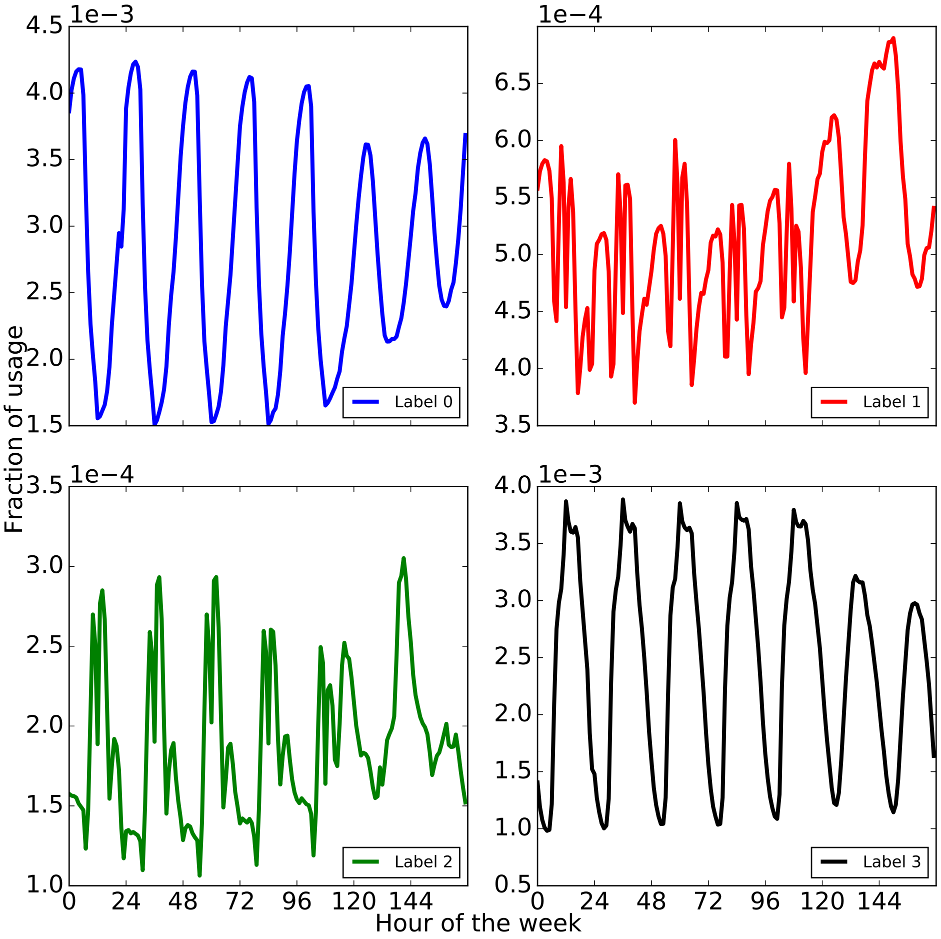

Figure 2.8: *Figure showing the percentage of usage for the labels 0 to 3 in the four labels mobility dataset for a weekly resolution. The figure resembles the behaviour seen in Figure 2.7, which also shows that the most frequent labels are labels 0 and 3. It can also be seen that the second most visited location, that is label 1, is mostly visited in the weekends, which may be the university as students are the users in the dataset. Label 2 may indicate many things, but could be supermarket or spare time work locations. Finally, label 3 accounts for all other spatial behaviour dominantly in the afternoon hours.*

top three locations, are the most frequent locations of users. Furthermore, labels 1 and 2 yield much smaller frequency, that is a order of magnitude 10 smaller than labels 0 and 3 and they seem to be more unpredictable than labels 0 and 3. In Figure 2.8 a similar figure is shown, yielding the fraction of usage for users in a weekly resolution. From Figure 2.8 it can be seen that the label 0 may indicate the users' home locations, as users mostly use this location in the night times. Label 1 may suggest the label of the university location, as this label is the most visited in the weekends rather than anywhere else. For label 1 it is also rather hard to find a specific pattern, which is because students have very different study plans from semester to semester, and therefore a specific periodic pattern for this label does not exist. Also given from the figure, label 2 may suggest a supermarket location or a spare time work location, as users mostly are at label 2 in the midday hours of the weekdays and Saturdays (again only valid for students). Finally, label 3 accounts for any behaviour that is not accounted for in the other discussed labels, which is mostly in the afternoon hours. The analyses performed on these labels are described in Chapter 6.

2.5 Combined social and spatial labels

This dataset is a dataset consisting of *artificial* labels. It is a four label dataset and consists of a combination of social and spatial labels. The dataset will be described in Section 7.1, since it is based on a combined social dataset, the **dialogue** dataset, which is introduced in Section 5.1.

Chapter 3

Methods

In this chapter, an overview of the methods used in this project will be given. First, an attack model will be presented in Section 3.1, which will discuss how users are uniquely identified in a large dataset. Next, the theory of Non-negative Matrix Factorization will be explained in Section 3.2, classification methods in Section 3.3, likelihood functions in Section 3.4 and temporal correlations in Section 3.5.

3.1 Attack model

In this section the attack model used to uniquely identify users will be discussed.
The attack model simulates the knowledge of a user by extracting random points in the dataset and then discarding the other users that are not equal to the user which is intended to be identified. The information granted from the random time bins are used as knowledge of a user, which otherwise could be acquired for example from social media (Facebook statuses, geo-tagged tweets, etc.) and can then be used to identify a person in the dataset. When the person is identified, the attacker knows who the given user is even though the dataset has been completely anonymized and therefore knows everything else about that user in the specific dataset, which is definitely a security breach. This is the concept of *unicity*. It should be mentioned that the terms *unicity* and *uniqueness* are used interchangeably in this thesis.

A problem when identifying a user may occur however if the user does not have much information in the database, either from missing data which could not be obtained at the given time, or if a user only contributed a limited amount of time to a study. Furthermore, if a user has missing data in a time bin in the database where the attacker knows the *real* value of the user, the information is useless to the attacker. To find out how the missing data impacts the unicity of a user in a large dataset, different scenarios for finding unicity will be presented. However, the general attack model for finding unicity will first be presented. In words, it is given as

1. Take a user from the dataset
2. Find a random time bin and extract the value for the given user from step 1. This step simulates a known external information about a user at a given time bin, which can be used to identify the user
3. Remove other users who do not have the same value in the same bin as the user in step 1.
4. Repeat steps 2. - 3. until no other users are left
5. Repeat steps 1. - 4. for all users

This model provides the attacker to easily identify users, as well as handling missing data, which otherwise would not be possible. The model can also be used regardless of the resolution (spatial or temporal) of the data. The model differs slightly given the different scenarios, which is given as pseudo-code in Appendix A for each scenario. The implementation of the model is slightly different, as it is ineffective

to find and discard users for each bin in each iteration. All users' values in each time bin have been found beforehand, of which one can then look up which users to keep. The implementation of this method used in this thesis can be seen in Appendix C.1.
Before stating the different scenarios, some notations will be given

$\boldsymbol{I}$ the index user who is intended to be identified

$\boldsymbol{D}$ the set of anonymized users in the original database

$\boldsymbol{d}$ an element in $\boldsymbol{D}$

$\boldsymbol{S}(\boldsymbol{k}) \subset \boldsymbol{D}$ the set of users who have the same trace as $\boldsymbol{I}$ after $\boldsymbol{k}$ points used

Here a trace is defined as the whereabouts of a user, that is a collection of spatio-temporal points. For example, if user 1 has the been the same places as user 2 at two different times, user 2 is still in S when $k = 2$. However, if user 3 has only been one of the places as user 1 and 2 at the same given time stamps, then user 3 is removed from S [14].

Two different types of access to the database are given, namely **direct access** and **indirect access**. Direct access means that if there is missing data then there is no event, where indirect access means that there can be an event, but it cannot be seen what the event is. Given these definitions, there are four different scenarios

Scenario 1 Direct access to I and direct access to D. When missing data for I is given, it should be defined as a new and well-defined label. When missing data for D is given, keep d in S if I has missing data for the time bin.

Scenario 2 Indirect access to I and indirect access to D. When encountering missing data for I, move on to the next label and do not increment *k*. When encountering missing data for D, keep d in S.

Scenario 3 Indirect access to I and direct access to D. When encountering missing data for I, move on to the next label and do not increment *k*. When encountering missing data for D, remove d from S.

Scenario 4 Direct access to I and indirect access to D. This scenario is the same as Scenario 2, but there is no missing data for I.

Scenarios 1-3 are the most interesting, which therefore are the ones discussed in this thesis. This is also because Scenario 4 is not well-defined, since no user has 100 % available data. Of the former three scenarios, the most probable scenario is Scenario 2, since missing data in this connection means that nothing has been observed, but there still is a true underlying value. This is because users always are *somewhere*, and it therefore does not make sense that there is no event at all. Pseudo-code for finding the uniqueness of a user for Scenario 1 can be seen in Algorithm 1, for Scenario 2 in Algorithm 2 and for Scenario 3 in Algorithm 3 in Appendix A.

As mentioned, the external information that is used to identify users in this project are just given from random time bins for each user. Therefore, one can *get lucky* and use only a couple of points to identify the user completely, where another time it may take a lot more points to identify the user. With this in mind, a definition will be given of the *95th percentile*, which is given in Definition 1.

Definition 1 (95th percentile)**.** *The 95th percentile is defined as the minimum number of points needed to identify a person 95 % of the time.*

For example, if the algorithm has been used for 1000 iterations, then the minimum number of points is given when the user has been identified 950 times. With the definition of the 95th percentile, a common measure for all users of their *findability* is given. This measure will be used for the different scenarios to see the importance of data availability. Additionally, for all datasets 1000 iterations has been used in the attack model.

3.2 Non-negative Matrix Factorization

In this thesis, Non-negative Matrix Factorization (NMF) will be used to find underlying structures of users' probability of being at home during the hour of the day or the hour of the week and will be used on the **Home/Work** dataset. The components will show how users behave and divide them into groups that have different behaviour of mobility.

NMF is a factorization or decomposition method that separates a matrix into components which can detect hidden structure in the original matrix. NMF is used as an unsupervised learning method, that is there is no given labelling or grouping of the data prior to the analysis. Given a data matrix $\mathbf{V}^{N \times p}$ with the constraint that all elements are non-negative, one can approximately represent it by decomposing it into two smaller matrices, that is

$$\mathbf{WH} \approx \mathbf{V} \tag{3.1}$$

where $\mathbf{W}^{N \times r}$ and $\mathbf{H}^{r \times p}$, $r \leq \max(N, p)$. The number of components r is defined beforehand, similar to the *K-means* clustering algorithm, which also needs the number of clusters before the algorithm is started. **W**, also called the *feature matrix*, holds the features in the columns. **H**, also called the *weight matrix*, holds the hidden structure found by the algorithm, and each element of **H** holds a weight for a specific feature in **W**. The elements of **W** and **H** are also all non-negative, and the matrices are found by maximizing the following function [24]

$$L(\mathbf{W}, \mathbf{H}) = \sum_{i=j}^{N} \sum_{j=1}^{p} [v_{ij} \log(\mathbf{WH})_{ij} - (\mathbf{WH})_{ij}] \tag{3.2}$$

The function given in (3.2) is the likelihood function based on the assumption that v_{ij} follows a Poisson distribution with mean $(\mathbf{WH})_{ij}$. Since the original paper that proposed the likelihood function in (3.2), other methods for finding the matrices have been proposed, and the implementation by the Python module *Sci-kit Learn* is defined as follows [25]

$$\arg\min_{W,H} \frac{1}{2} \|\mathbf{V} - \mathbf{WH}\|_{Fro}^{2} = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{p} [v_{ij} - (\mathbf{WH})_{ij}]^2 \tag{3.3}$$

NMF is ideal for analysing images, which all have non-negative pixel values or for clustering corpora of articles into topics based on TF-IDF values of words in the articles. Another application is to lower the dimensions of the data for example for visualization purposes, similar to *Principal Component Analysis*. A flaw of the NMF however, is that the representation of the decomposition may not be unique, meaning there is not one unique representation of **W** and **H** [26, Ch. 14]. A NMF implementation can be found in the *Sci-Kit Learn* module for the Python programming language [25].

3.3 Classification methods

To see if users can be classified based on the features found from Non-negative Matrix Factorization (see Section 3.2), classification methods will be used to determine how informative these features are. The classification methods used are

- ◇ K-Nearest Neighbor (KNN) classifier
- ◇ Random Forest (RF) classifier

The KNN classifier is a simple classifier that finds the *K* nearest neighbors of a point in space and simply assigns a class to the point based on a majority vote of the classes of the neighbors. If the vote is tied, a random class is assigned of the tied classes.
The RF classifier is an *ensemble method*, meaning that it combines a lot of other classifiers to improve performance. The RF classifier specifically combines *decision tree classifiers*, hence the name *random forest*. Decision trees will not be discussed in detail as the focus mainly is to compare the performance of the RF classifier to the performance of the KNN classifier [27, Ch. 5].
To find the optimal parameters in the classifiers, K-fold Cross-Validation methods will be used. Furthermore, after determining the optimal parameters, the dataset will be split into a training set and a test set to determine the performance of the classifiers [28]. The KNN and Random Forest classifiers are both found in the *Sci-Kit Learn* module for Python [29, 30].

3.4 Likelihood

The idea of using likelihood estimations can show which users are more likely to follow a certain distribution than others. This means that the users that have the largest likelihood value will be more similar to other users and should in turn show which users are more likely to be found than others. This method will be used on the **Home/Work** dataset.
The likelihood function of a user is defined as the product of the probability density functions (PDF) for each time t , that is

$$L_u = \prod_t f_{u,t} \tag{3.4}$$

where L_u is the likelihood of a given user, t is the time (which can be between 0 and 23 for daily hours or 0 and 167 for weekly hours) and $f_{u,t}$ is the probability density function. The PDF should be well-defined and found from an already known distribution, where the distribution in this case is found from the fraction of ones at a given time for all users in the defined resolution. This could for example be the binomial distribution. However, a well-known PDF is not well suited for this problem, as the data in the **Home/Work** dataset does not resemble any well-known distribution, so the raw values of the distributions will be used instead. This means that the distributions are binned and $f_{u,t}$ are found as the value of a user for the given bin. The values have been weighted such that the integral over the range equals 1.

Usually, the *log-likelihood* is preferred, since the likelihoods can be very small and hard to inspect. It can also help the computation to take the log-likelihood instead, and the product of the PDF's becomes a sum. The log-likehood is then defined as

$$\log L_u = \sum_t \log f_{u,t} \tag{3.5}$$

where $\log$ is the natural logarithm.

3.5 Temporal correlations

The method of finding uniqueness (see Section 3.1) may be influenced by which time bins are used to identify a user. An analysis regarding the importance of time bins when identifying a user will therefore be conducted, which shows which time bins are the most important if a user needs to be identified. This means that for example a time slot in the night may have more information than one in the middle of the day, and therefore it is more desirable to know something about a user in the night such that he or she may be easier to identify.

To find out *when to ask* about a user or have information about a user given a specific time, the following approach has been taken: For each identification process of a user (index user), each time a time bin has been selected, remove the users who are not equal to the index user. Then, divide the number of users equal to the index user with the number of users equal to the index user at the previous time bin. In other words, if $|S(k)|$ is the number of users left after removing users not equal to the index user after using k time bins, the fraction $\frac{|S(k+1)|}{|S(k)|}$ is saved in the given resolution represented by the time stamp at point $k+1$. Therefore, a value of 1 means that no users were discarded at point $k+1$ and a value of 0 means that all users were discarded at point $k+1$. The resolution is chosen to be either hour of the day or hour of the week. This approach will be used for different groups of users, that is for all users and the users who have the highest, lowest and median 95th percentiles for all datasets. The results found from this analysis can be used to clarify at which point in time the most valuable information can be found for different kinds of users and in the different datasets when identifying users.

If the dataset consists of two labels, the *density* can be compared to the temporal correlations. The density, denoted by ρ, is the global probability of observing a 1 at a given time bin, that is the fraction of 1's for a given time bin. If D is the set of users with different labels than the index user, then

$$|D| = P(\text{user has label } 0) \cdot n(1) + P(\text{user has label } 1) \cdot n(0) \tag{3.6}$$

where P is the probability and n is the number of users having label 0 or 1. Since $n(0) = N \cdot (1-\rho)$ and $n(1) = N \cdot \rho$, where N is the total number of users and $P(\text{user has label } 1) = \rho$, then

$$\begin{aligned} |D| &= (1-\rho) \cdot N \cdot \rho + \rho \cdot N \cdot (1-\rho) \\ &= 2 \cdot N \cdot \rho \cdot (1-\rho) \end{aligned} \tag{3.7}$$

From (3.7) we get that the fraction of users to discard at any time bin on average, is proportional to $\rho \cdot (1-\rho)$. This is given if one assumes that users are random and independent, which is a valid assumption. It is therefore intended to compare the results in (3.7) with the results given from the temporal correlation analysis of datasets having two labels. The curve $\rho \cdot (1-\rho)$ denotes the fraction of users to discard so the curve $1 - \rho \cdot (1-\rho)$, denoted by the *density curve*, will be compared to the temporal correlations. The inclusion of the density curve will be given for the two datasets that consist of two labels, that is the **Home/Work** dataset in Section 4.5 and the social labels dataset in Section 5.3.

Chapter 4

Home/Work labels

In this chapter an analysis of the **Home/Work** dataset will be discussed. The attack model as well as the other methods described in Chapter 3 will be used to measure the findability of users in the dataset.

4.1 Unicity

In this section the results given by applying the attack model in Section 3.1 on the **Home/Work** dataset will be given. For all scenarios described in the model, boxplots have been made for users with different data quality, that is three for each of three groups of quality. The different qualities are defined as *good* if a user has a low fraction of missing data, that is between 0 and 40 %, *medium* if between 40 and 60 %, which is a medium fraction and *bad* if the fraction is high, that is between 60 and 100 %. Furthermore, the 95th percentile for all users has also been calculated, as defined in Definition 1. For all scenarios, the results shown are found with 1000 iterations to reduce randomness, since random bins are used each time a user should be identified. It should also be mentioned that a maximum number of points has been chosen to 100. This is because if the user has not been uniquely identified after 100 points, then the user is very hard to identify and using more points is therefore irrelevant. Doing this will also lower the computation time accordingly.

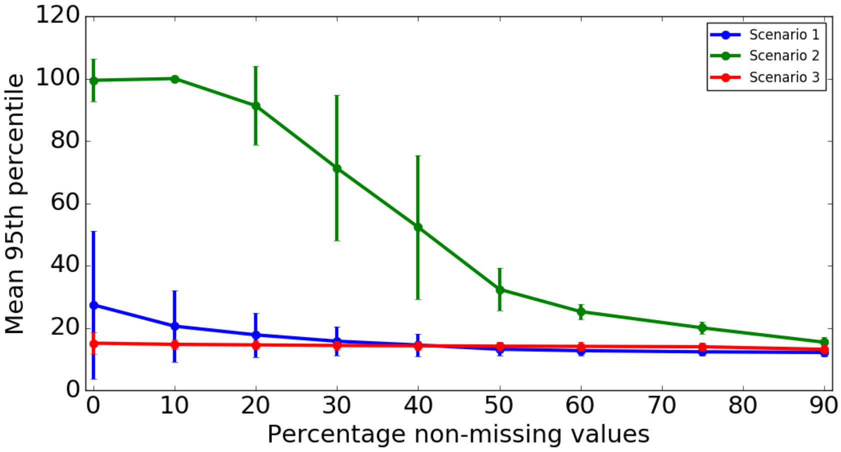

Figure 4.1: *Graph showing the mean 95th percentile as a function of the percentage of non-missing data of users. The errorbars show the standard deviation of the distributions of the 95th percentiles. It can be seen that many users need to be removed in Scenario 2 to get decent results, and that Scenario 3 has a constant mean 95th percentile no matter how many users are present. When removing users in Scenario 1 the mean quickly converges. It can be seen that the standard deviation of the distributions in Scenario 2 becomes larger when removing more users, and eventually decreases when moving from 40 to 50 % non-missing data.*

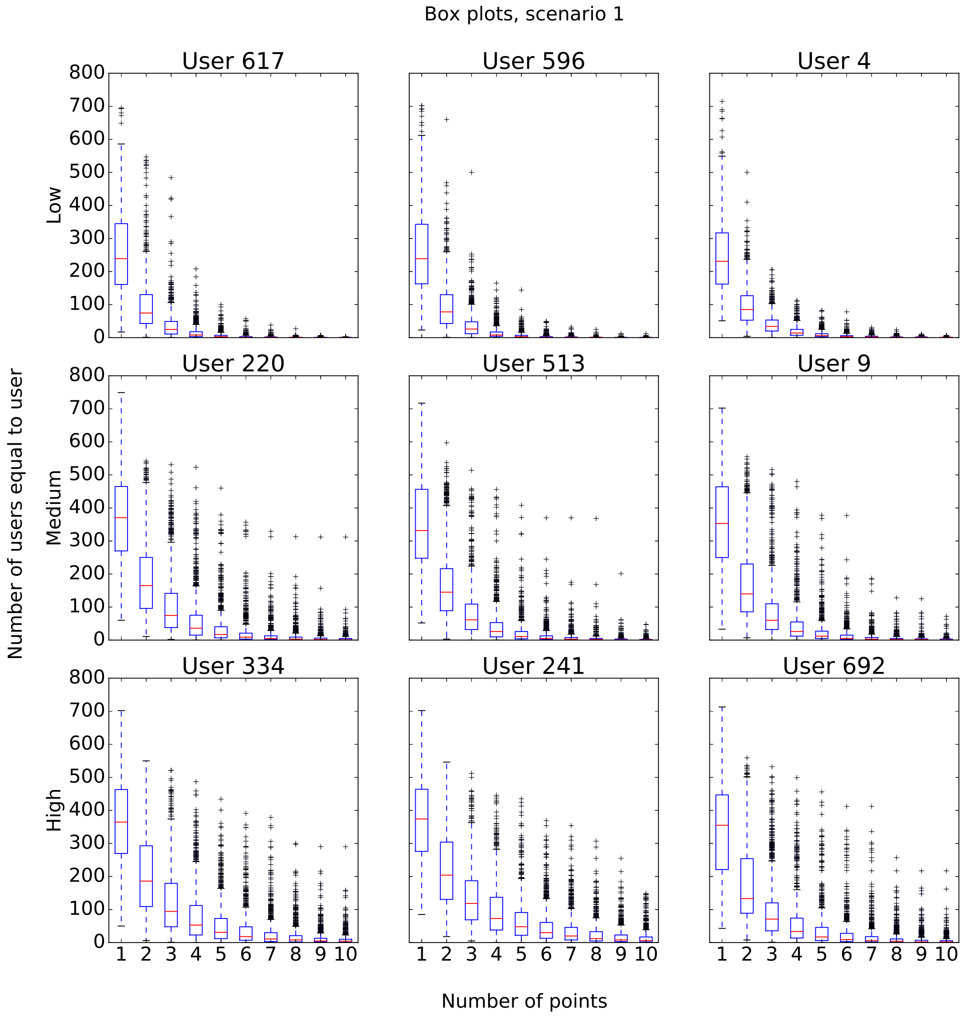

Figure 4.2: *Boxplots for Scenario 1 for nine different users having different data quality. Low means good quality, medium is medium quality and high means bad quality. The box plots show the convergence for the first 10 points used. It can be seen that the data quality has an impact on the uniqueness of a user, where bad quality yields a harder identification and good quality yields an easy identification.*

Before the analyses can be performed some pre-processing of the data needs to be performed. This is because some users may not be identified even after using 100 points as external information. This observation is evident in Figure 4.1. In Figure 4.1 the mean 95th percentile of all users is shown as a function of percentage of non-missing values for users. In other words, users who have less than some percent non-missing data have been removed, and the mean of the 95th percentiles has been calculated for the remaining users. Additionally, the figure shows the standard deviation of the distributions of 95th percentiles for the scenarios as errorbars.

From Figure 4.1 one can see that pre-processing of data for Scenario 2 is needed for further calculations, since only a few users are uniquely identified when considering all users. This can be seen because the graph has reached a value of close to 100 on the second axis and the distribution is quite narrow. The reason it is not zero is because a few users have been identified and therefore lowers the mean and increases the standard deviation accordingly. Therefore, the computation has stopped for most of the users when not removing any users for Scenario 2, since the maximum number of points is 100. This is also evident when only removing users with less than 10 % valid data, where even fewer users have been

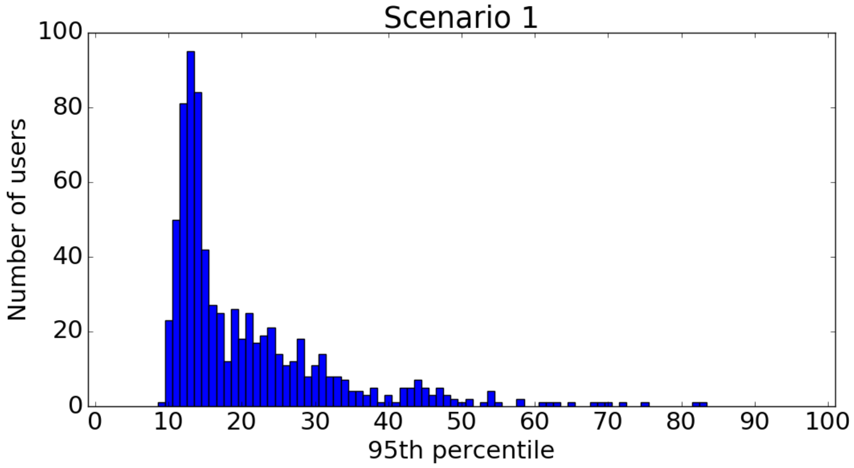

Figure 4.3: *Histogram showing the 95th percentiles for users who have more than 10 % non-missing values for Scenario 1. All users are identified, however some users need more points when identifying than others.*

identified. It can also be seen that the mean is decreasing for Scenario 1, where it seems to be almost constant for Scenario 3. For Scenario 1, the standard deviation becomes much smaller when removing users with less than 10 % non-missing data, but for Scenario 3 it is almost constant. In Appendix B.1 one can see the number of users in the dataset given different percentages of valid data of users.

To get decent results without removing too many users, users with less than 10 % non-missing data have been removed for Scenario 1. Removing these users yield 755, having removed 80 users with bad data quality. The percentage is based on Figure 4.1, where the standard deviation becomes much smaller than when considering all users. Additionally, this is the lowest percentage needed to uniquely identify all users. In Figure 4.2 one can see box plots calculated for nine users having different data quality in Scenario 1. The box plots show the number of users equal to the user which is being identified after *k* points. Looking at Figure 4.2, it is clear that each user can be uniquely identified with a low number of points, and that users who have a good data quality can be found with a lower number of points compared to users having a bad data quality.

In Figure 4.3 one can see a histogram showing the distribution of the 95th percentiles. When inspecting Figure 4.3, one can see that the distribution is skewed with a large number of users having their 95th percentile around 14. Increasing the number of points after 14 sharply decreases the number of users, while some still need a lot of points to be identified. This may be caused by their data quality, as Figure 4.2 shows that the data quality has an impact on the 95th percentile. This is because missing data is treated as a new label, and if users have many missing data points they can be hard to distinguish.

Regarding Scenario 2, some more processing of the data needs to be done, otherwise no users can be identified. This is caused by the fact that when meeting missing data for other users, one should skip the time bin and move on to the next. Since many users have bad data quality, this causes the algorithm to never uniquely identify any user. Therefore, users with less than 50 % non-missing data have been discarded, which is based on Figure 4.1. This percentage is chosen since there is a large drop in the curve in Figure 4.1, as well as the standard deviation decreasing much from considering users with less than 40 % non-missing data. This means that all of the users who have bad data quality have been removed, as well as some of the users who have medium quality. After removing the users 413 remain, that is 422 users has been discarded.

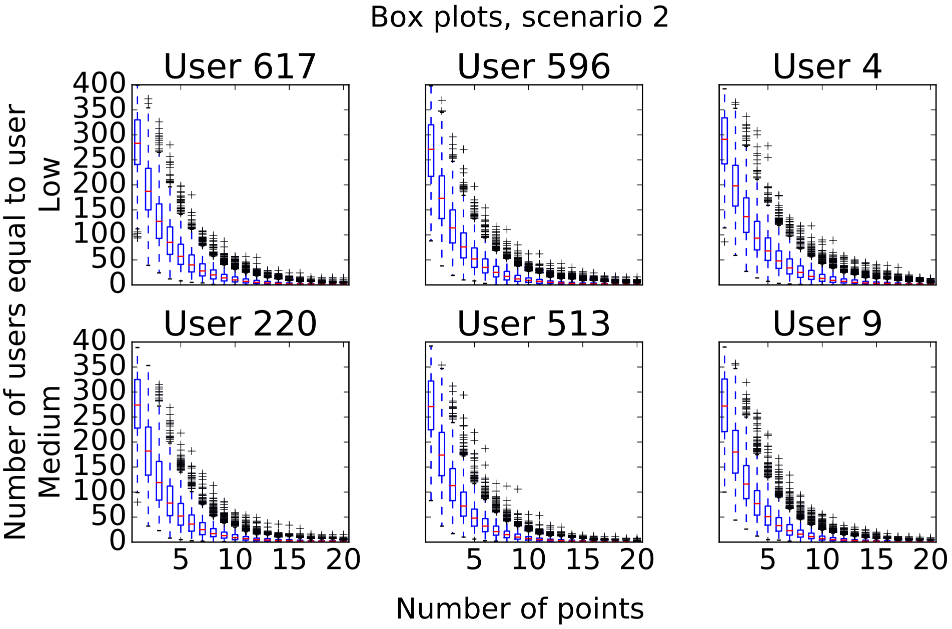

Figure 4.4: *Boxplots for Scenario 2 for six different users having different data quality. Low means good quality and medium is medium quality. The box plots show the convergence for the first 20 points used. It seems as though the data quality does not have much importance as opposed to Scenario 1, which can be seen in Figure 4.2.*

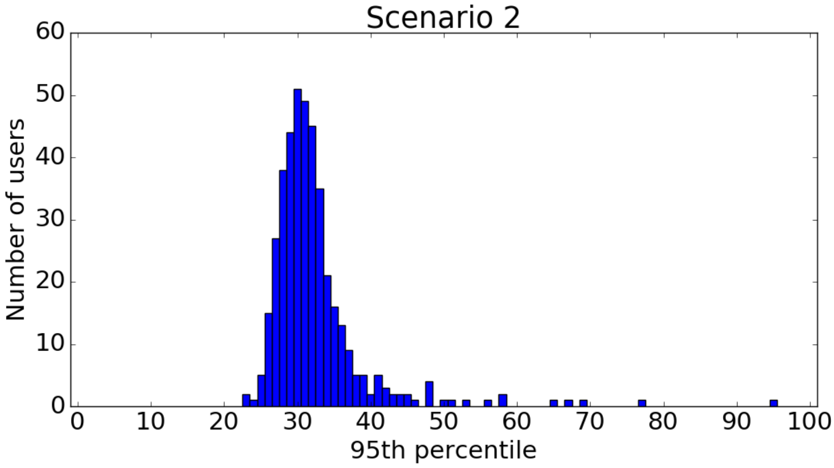

Figure 4.5: *Histogram showing the 95th percentile for users who have more than 50 % non-missing values for Scenario 2. All users are identified, but with the expense of removing more than half of the total amount of users. In general, the number of points needed to identify a user is quite high, as opposed to Scenario 1.*

In Figure 4.4 one can see the box plots of six users, three having good data quality and three having medium data quality. From Figure 4.4 it can be seen that the convergence towards uniqueness is slower for Scenario 2 than Scenario 1, meaning that in general more points are needed to uniquely identify users. It also seems that the difference between good data quality and medium data quality is not that important in Scenario 2. This could mean that Scenario 2 treats all users in the same way, even though they have missing data, which is important. Scenario 2 is also the most probable, since as mentioned

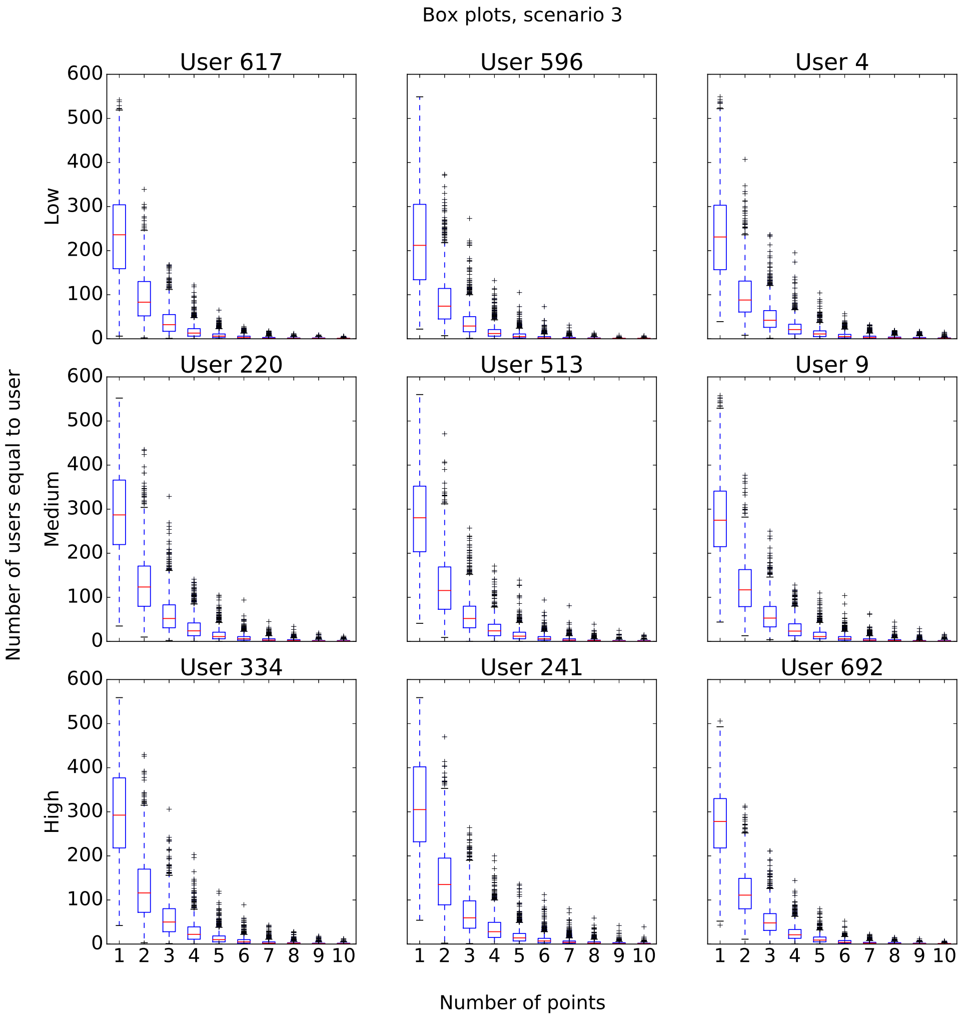

Figure 4.6: *Boxplots for Scenario 3 for nine different users having different data quality. Low means good quality, medium is medium quality and high means bad quality. The box plots show the convergence for the first 10 points used. From the figure it seems as though the data quality does not have much importance when identifying users, and the convergence is quite fast, as opposed to Scenarios 1 and 2, which can be seen in Figures 4.2 and 4.4.*

in Section 3.1 users are always *somewhere*, and that missing data should mean that there is an event, but that one cannot see what it is. In Figure 4.5 one can see the 95th percentile histogram for Scenario 2. Looking at Figure 4.5 it can be seen that the distribution is almost normal, but with a tail occurring after 45 points. Also, all users have been identified, but with the expense of removing more than half. This means that for Scenario 2, it is evident that users need good data quality to uniquely identify them. Furthermore, the mean of the distribution is higher than for Scenario 1, which also is seen in Figure 4.1.

For Scenario 3, users with less than 10 % non-missing data have been removed. This percentage has been chosen to remove those users who are very hard to find and indeed have too much missing data so it does not make sense to include them. In Figure 4.6 nine box plots are given for users with different data quality for Scenario 3. From Figure 4.6 it can be seen that data quality does not have a great impact in the way of calculating uniqueness for Scenario 3. This is caused by the fact that when encountering missing data for other users than the index user, they are discarded. Therefore, one assumes that they do not offer any extra information, which is not the case in the real world, as users as said always are *some-*

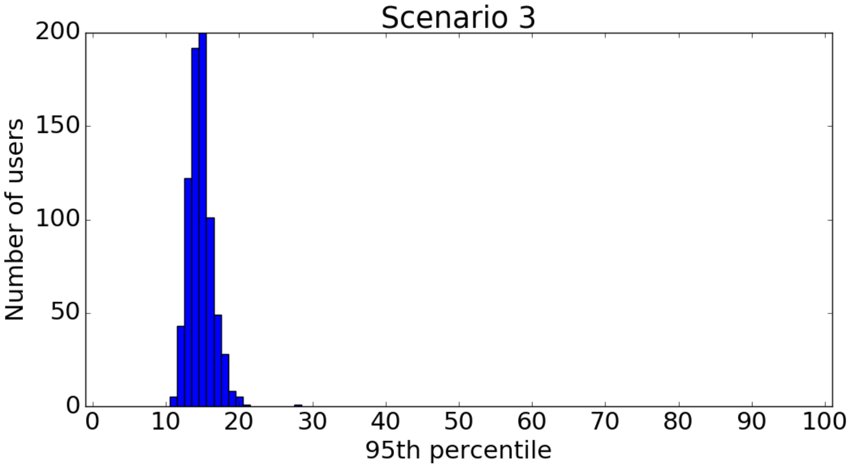

Figure 4.7: *Histogram showing the 95th percentile for all users for Scenario 3. Users are generally easier to find in Scenario 3 than in Scenarios 1 and 2, and the 95th percentile almost seems normally distributed. The distribution is much more narrow than the distributions for Scenarios 1 and 2 seen in Figures 4.3 and 4.5. However, users' missing data are ignored, where the missing data might have importance.*

where. It can also be seen from Figure 4.6 that in general a few number of points is needed to uniquely identify users. In Figure 4.7 one can see a histogram of the 95th percentiles for Scenario 3. It can be seen from Figure 4.7 that the distribution seems normal, with a single outlier at 28 points, and is a much more narrow distribution than for scenarios 1 and 2. Users are therefore in general easier to identify in Scenario 3 than scenarios 1 and 2. These results are interesting, but Scenario 3 is as mentioned not that relevant regarding the real world. Users can still be identified however and an attacker can still breach the privacy of a user, so the results are still important to highlight.

To see the individual spread of identifications within users having different findability, a figure showing histograms for different users in all scenarios is seen in Figure 4.8. In Figure 4.8 one can get a better understanding of the 95th percentiles. It can be seen that 95 % of the identifications are to the left of the dashed lines. Therefore, as the findability increases, the dashed lines move further to the right on the first axis. It can also be seen that each scenario is different, as also seen when comparing the histograms in Figures 4.3, 4.5 and 4.7. For users with low findability, that is a high 95th percentile, there is a large spread in the distributions and users 739 and 813 have not been identified in all iterations. This is seen by the number of iterations at 100 points, since the algorithm stops when this number of points has been reached. As the findability increases, the distributions become more narrow. Therefore, users who are easy to find are almost always easy to find and will eventually be identified.

In Figure 4.9, one can see a figure that is similar to the figures given by de Montjoye et al. [14, 15]. The figure shows the mean fraction of unique users after each number of spatio-temporal points considered over all iterations, which de Montjoye et al. expresses as $|S(I_p)| = 1$ meaning that only one user has the exact trace of points sampled. The authors also consider the fraction of users which have the feature $|S(I_p)| \leq 2$, but this will not be considered since the figure is only a measure of comparison. It is given in [14] that four spatio-temporal points are enough to uniquely identify 95 % of all users. Looking at Figure 4.9, one can see that the lowest number of points needed to uniquely identify 95 % of all users is 15 points for Scenario 3. This is much more than the results given in the article, so this indicates that using **Home/Work** labels instead of explicit (longitude, latitude) coordinates is more privacy preserving.
Additionally in Figure 4.9, the mean 95th percentile has been found for all scenarios as well as the

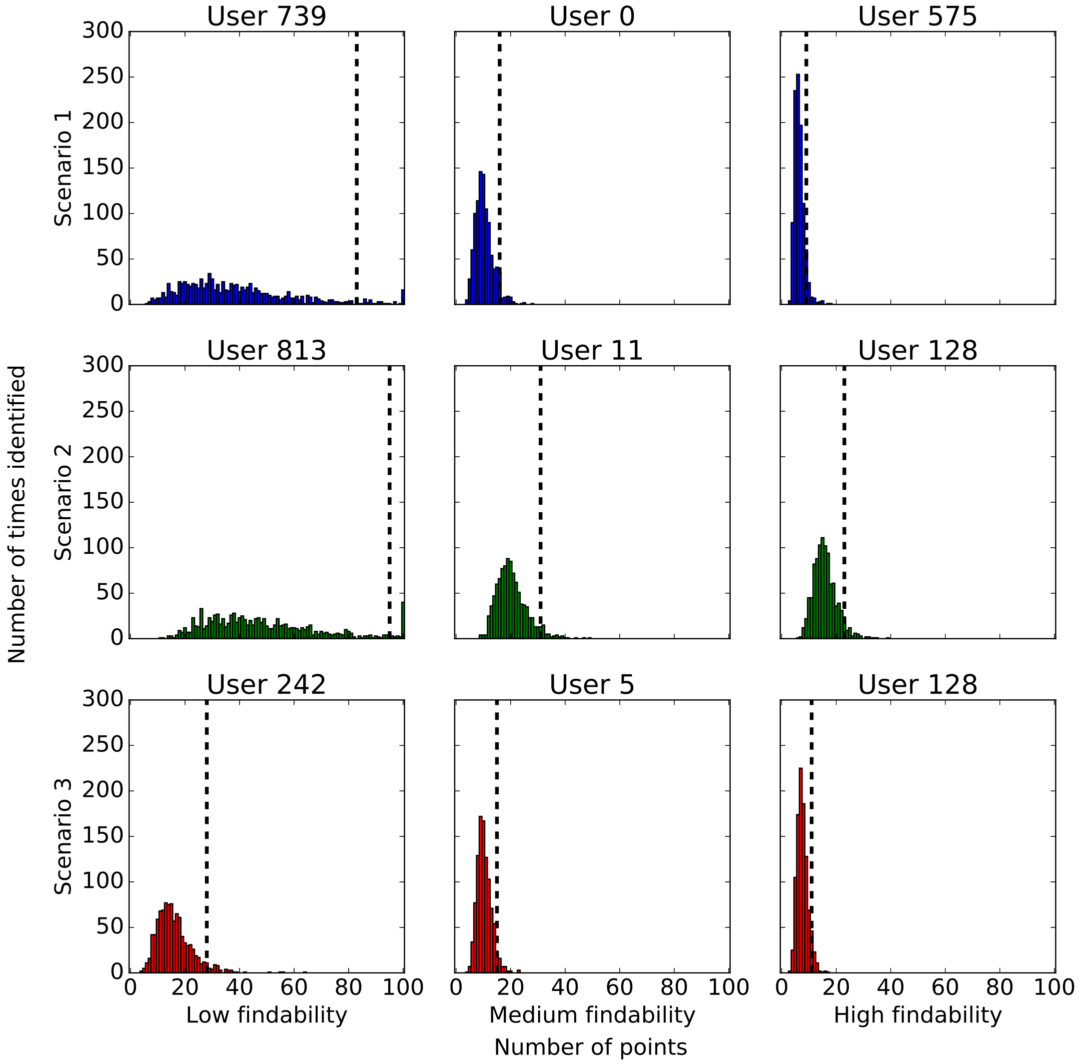

Figure 4.8: *Histograms showing the number of identifications for different types of users. For users with low findability, users with the maximum 95th percentile has been chosen, where users with the median 95th percentile has the medium findability and users with the minimum 95th percentile has high findability. The dashed lines indicate where the 95th percentiles are placed for users. It can be seen that for users 739 and 813 some iterations have not uniquely identified them, as there are some iterations that have used 100 points before stopping the algorithm. The higher the findability the more narrow the distributions are, meaning users who are easily identifiable have a tendency to always being able to be identified no matter which time bins are considered.*

standard deviation shown as green and red bars, respectively. This is done such that the method by de Montjoye et al. can be compared to the 95th percentiles to see if the 95th percentiles are valid as a comparison measure. The 95th percentiles show an individual measure of findability, so if the results are equivalent the 95th percentiles are to be preferred rather than just aggregating over users. From Figure 4.9 it can be seen that more points are needed to identify users in Scenario 2 than the other scenarios, where Scenario 3 needs the least number of points. The mean 95th percentiles are close to the minimum number of points needed to identify 95 % of all users for Scenario 2 and 3, but not as close for Scenario 1. As the distribution of the 95th percentiles for Scenario 1, as seen in Figure 4.3, has a long tail and is not completely normally distributed, this may cause the mean to be off. It can also be seen

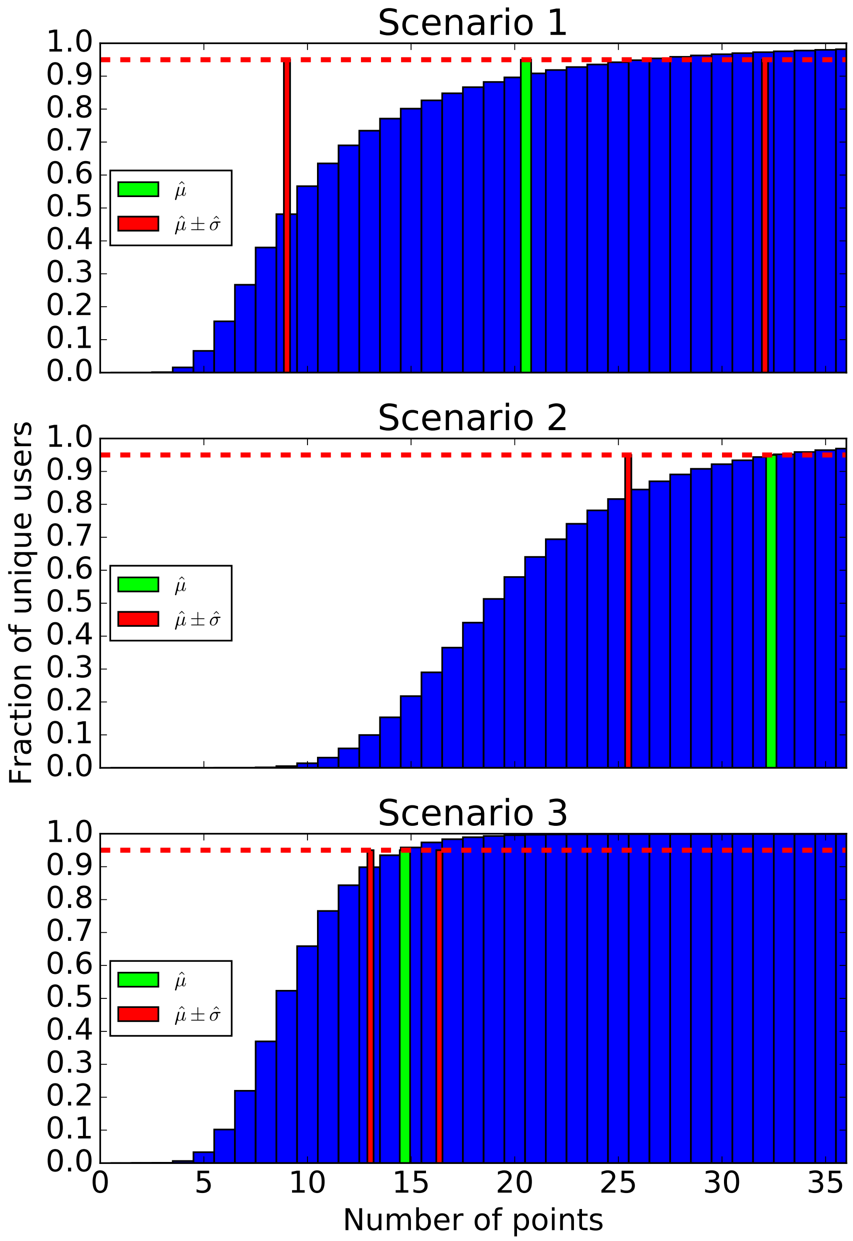

Figure 4.9: *Figure showing the mean fraction of users identified at each number of points for the three scenarios. The green bars show the mean 95th percentiles and the red bars show the spread of the standard deviation from the mean. It can be seen that taking the mean of the 95th percentiles almost equals the mean fraction of unique users when looking at 95 % identified, however with some deviations for Scenario 1. The paper by de Montjoye et al. indicates that four spatio-temporal points are enough to uniquely identify 95 % of all users [14], where at least 15 points are needed for Scenario 3 when using* ***Home/Work*** *labels instead of explicit (longitude,latitude) coordinates.*

that the uncertainty expressed by the standard deviations is much larger for Scenario 1 than for the other scenarios. If more users with bad data quality are removed, the distribution of the 95th percentiles may be more normally distributed for Scenario 1, so by doing this the mean may be closer to the minimum number of points needed to identify at least 95 % of users. Users' individual behaviour are not clarified however when presenting the figures such as Figure 4.9, which is why the 95th percentile analysis has been performed.

From the results in this section, it is clear that when considering scenarios 1 and 3 it is easier to identify users than in Scenario 2, which is easily seen in Figure 4.1. To identify users in Scenario 2, a lot of users had to be removed, which ultimately led to the identification of all users so a trade-off must be given. Direct access given for Scenarios 1 and 3 however, as defined in Section 3.1, does not make much sense when looking at mobility data, so Scenario 2 is the most relevant when considering handling missing data. The results also indicate that for Scenarios 2 and 3 the data quality does not make much of a difference when identifying users, as seen in Figures 4.4 and 4.6, but for Scenario 1 it does as seen in Figure 4.2. Comparing the results in Figure 4.9 with the work done by de Montjoye et al. also indicates that when using **Home/Work** labels instead of full GPS-coordinates, users are harder to find since at least 15 points are needed in Scenario 3 to uniquely identify 95 % of all users, with even more points needed for scenarios 1 and 2. Furthermore, scenarios 2 and 3 show that when aggregating over users, one can still use the 95th percentiles, but a larger deviation occurs for Scenario 1. However, the 95th percentiles are still to be preferred, since these show the individual findability of users.

4.2 Non-negative Matrix Factorization

In this section the results from the Non-negative Matrix Factorization (NMF) analysis will be given. The theory of NMF is given in Section 3.2, and the terminology will therefore be the same in this section. Since the data matrix **V** must hold non-negative values, the elements of **V** will consist of fractions of 1's, that is the probability of being at home, for all users in a given resolution. This means that **V** will either have the dimension $\mathbf{V}^{N\times24}$ or $\mathbf{V}^{N\times168}$, where N is the number of users. N can vary since the users

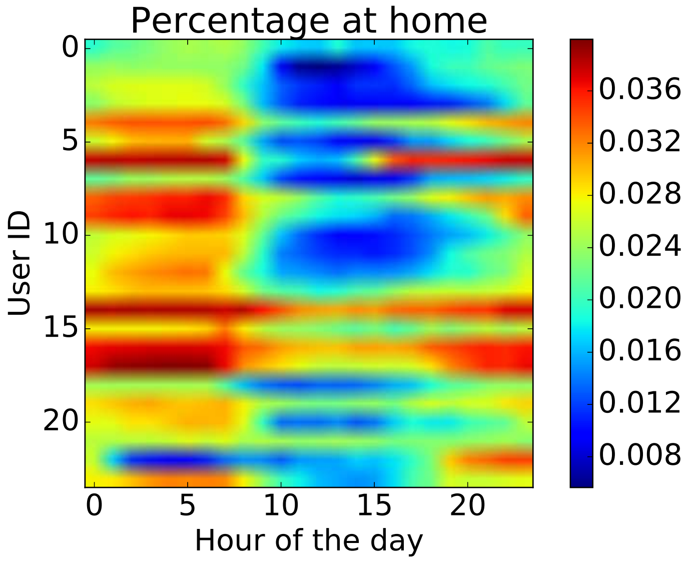

Figure 4.10: *Figure showing the daily percentage of being at home for the first 24 users/rows of* $\mathbf{V}^{835\times24}$. *It can be seen from the figure that users usually are home in the night-time and evenings and away from home in the middle of the day.*

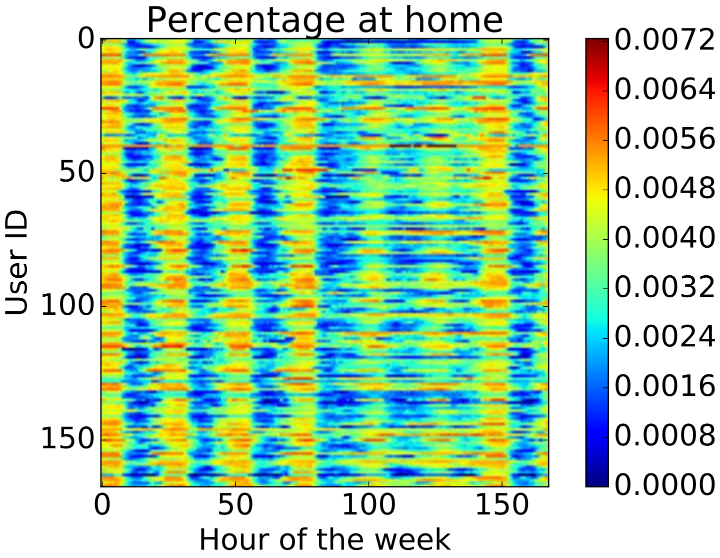

Figure 4.11: *Figure showing the weekly percentage of being at home for the first 168 users/rows of* $V^{835\times168}$. *It can be seen from the figure that users are highly regular in their whereabouts, with a very specific weekly pattern. Users are in general more at home on Saturdays and Sundays.*

with the worst data quality will be removed and the findability of users depends on the scenario chosen, which determines the number of users. It should be mentioned that the probability of being at home is calculated as number of 1's (that is the number of times the user was at home at a given hour of the day or hour of the week) divided by the number of available observations at the given time. Otherwise the probabilities would not represent the user's behaviour correctly, which only can be found from the available data. An illustration of the percentage at home for the first 24 users on a daily basis can be seen in Figure 4.10. Only the first 24 users have been shown such that the figure is squared.

From Figure 4.10 it can be seen that there is a general trend that users are away from home in the middle of the day and are at home in the evening and at night, which is expected. In Figure 4.11 one can see the behaviour of the first 168 users during the week. Again, the number of users has been chosen such that the figure is squared and easier to inspect. Looking at Figure 4.11 one can see a general trend which is similar to the one in Figure 4.10 and the day time can clearly be separated from the night time. It can also be seen that users are more at home on Saturdays and Sundays. However, it can also be seen that some users deviate from the general trend, which are the users that might be separated from the common users by the NMF. Both in Figure 4.10 and 4.11 all users have been kept, that is no users have been removed even though they may have bad quality. However, when making the NMF analysis, users with less than 10 % valid data have been removed, since these users will have a strange impact on the results and also because it generally makes no sense to include them. This will give a matrix $\mathbf{V}^{755\times24}$.

To make things simple, two components in the NMF has been chosen. This is because there may not be a large difference in users, and a simplification of two groups of users/trends of users will seem as a plausible choice when basing it solely on the probability of being home or not. Also, since NMF is an unsupervised learning method, there is generally no easy way to find the number of components, and therefore a simple choice of two components has been taken. First, an analysis based on the daily representation will be given and then a weekly representation will be discussed, starting with the former.

Daily NMF

In this section, the results from the daily resolution of $\mathbf{V}^{755\times 24}$ will be given. Running the algorithm yields a $\mathbf{W}^{755\times 2}$ matrix and a $\mathbf{H}^{2\times 24}$ matrix, where **W** holds the scores and **H** holds the components, respectively. Inspecting the components yields the plot seen in Figure 4.12. From Figure 4.12 it can be seen that the algorithm highlights two different kinds of users: Users who are mostly home at night, which can be seen from Component 0 and users who are mostly at home during the day, which can be seen from Component 1. This can be seen since a linear combination of the components with the values in the matrix **W** will give the matrix **V**, and the weights in **H** weights each user. Therefore, a user that has higher values in **V** that corresponds to the indices in **H** will be weighted more so. The components found from NMF can therefore be seen as a clustering property, where each component in theory should resemble a cluster. However, in this case the clusters will with high probability be skewed, since most users are at home during the night. Therefore, a simple clustering which assigns each user to the score in **W** that has the highest value has not been used, since as mentioned the number of users in each cluster would be skewed, with a large amount of users being assigned to one cluster and only a few to the other.

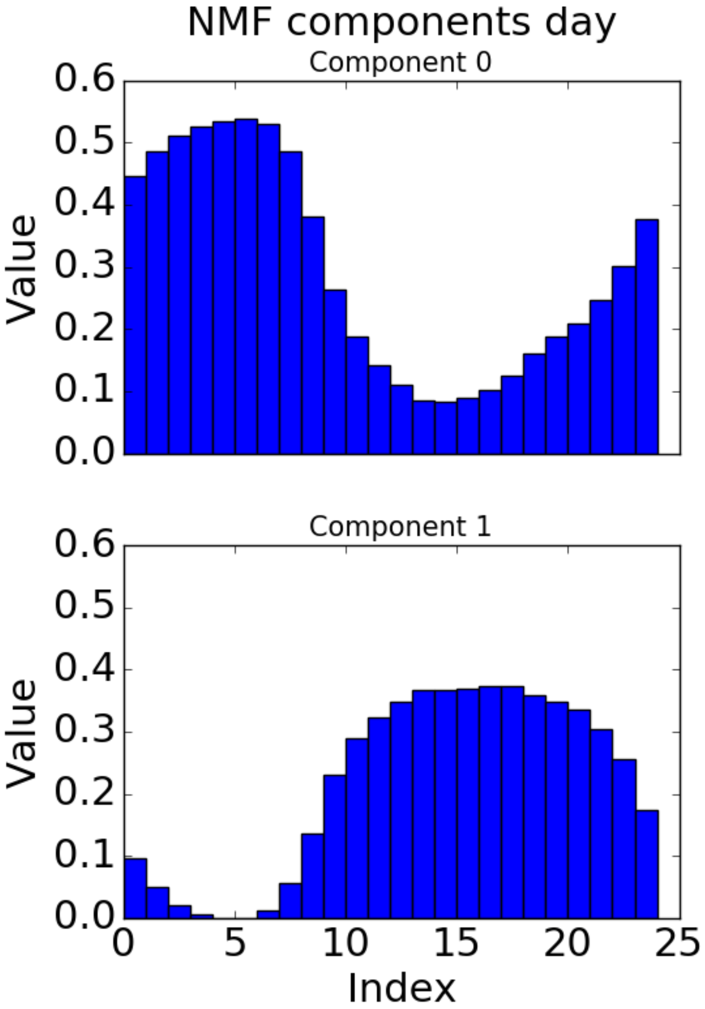

Figure 4.12: *The components of **H** for the daily representation of **V**. It can be seen that two different types of users can be found from the NMF: Users who are mostly home at night and users who are mostly home in the afternoon and evening hours. It is hypothesised that Component 0 highlights the general trend of users and Component 1 highlights a smaller group of users.*

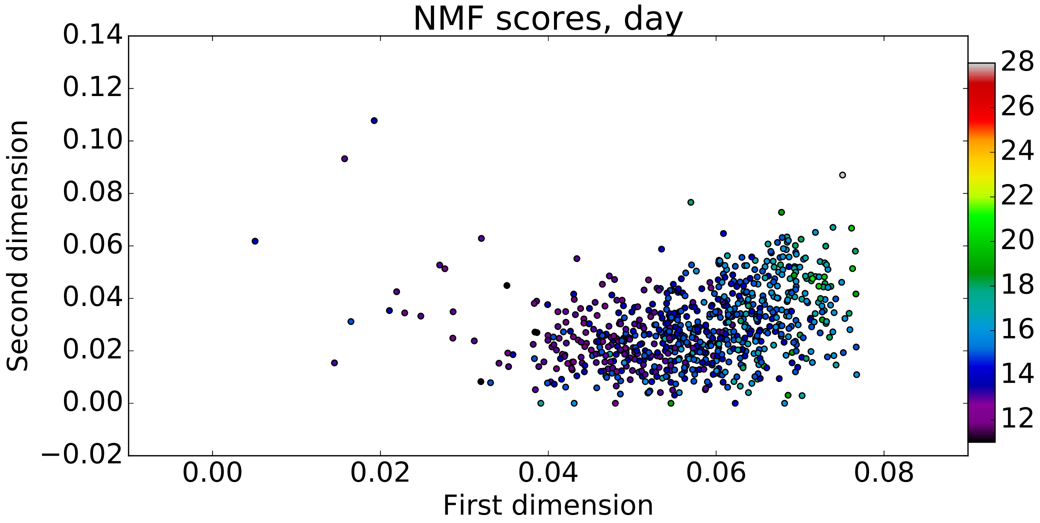

Figure 4.13: *The scores of the daily NMF. The colour of the points correspond to the 95th percentile found from Scenario 3. The plot suggests that the first dimension accounts for the findability of a user, that is a larger value in the first dimension means a larger value of the 95th percentile.*

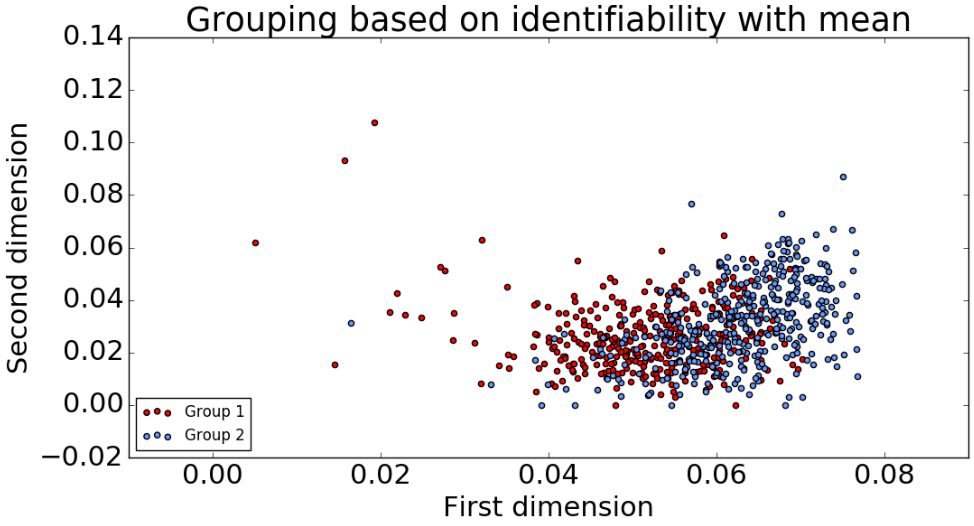

Figure 4.14: *Group 1 and group 2 coloured in NMF space for the daily representation. Group 1 contains the users who have less than or equal to the mean 95th percentile for Scenario 3, and group 2 contains the users who have larger. It can be seen that the most of the users in group 2 have a larger value than the users in group 1. It therefore seems that the users can be classified based on their NMF scores, which is done in Section 4.3.*

Plotting the scores of the NMF, that is the columns of **W** gives the plot seen in Figure 4.13. The colouring seen in the figure is the 95th percentile found from Scenario 3. Looking at Figure 4.13 one can see that there seems to be a trend that the higher the value in the first dimension (dimension 0), the higher the 95th percentile gets. That is, the users who are easy to find generally have a lower value in the first dimension rather than the second dimension. This observation makes sense, since the first dimension represents users who are mostly home at night, and since this is a common behaviour, the more a user resembles the average user (whom are mostly characterized by the first dimension), the harder it is to identify them.

Therefore, as the value of the first dimension is increased, the 95th percentile also increases. To verify this, two groups of users has been made: users who have less than or equal to the mean 95th percentile and users who have a value that is higher than the mean 95th percentile. These two groupings are called **group 1** and **group 2**, respectively. Group 1 has 362 users and group 2 has 393 users, so the groups are close to being evenly distributed. From this grouping, a plot has been produced which colours the groups in the NMF space, which can be seen in Figure 4.14. It can be seen in Figure 4.14 that the groups are not completely separated, however it can be seen that indeed an increase in the first dimension yields a transition from group 1 to group 2. To see how much importance the NMF scores have on the groups, a classification of the two groups based on their NMF scores will be performed in Section 4.3.

Weekly NMF

To see if the resolution has an impact on the NMF results, an analysis based on the weekly representation of $\mathbf{V}^{755\times168}$ has also been made. The decomposition yields the matrices $\mathbf{W}^{755\times2}$ and $\mathbf{H}^{2\times168}$. In Figure

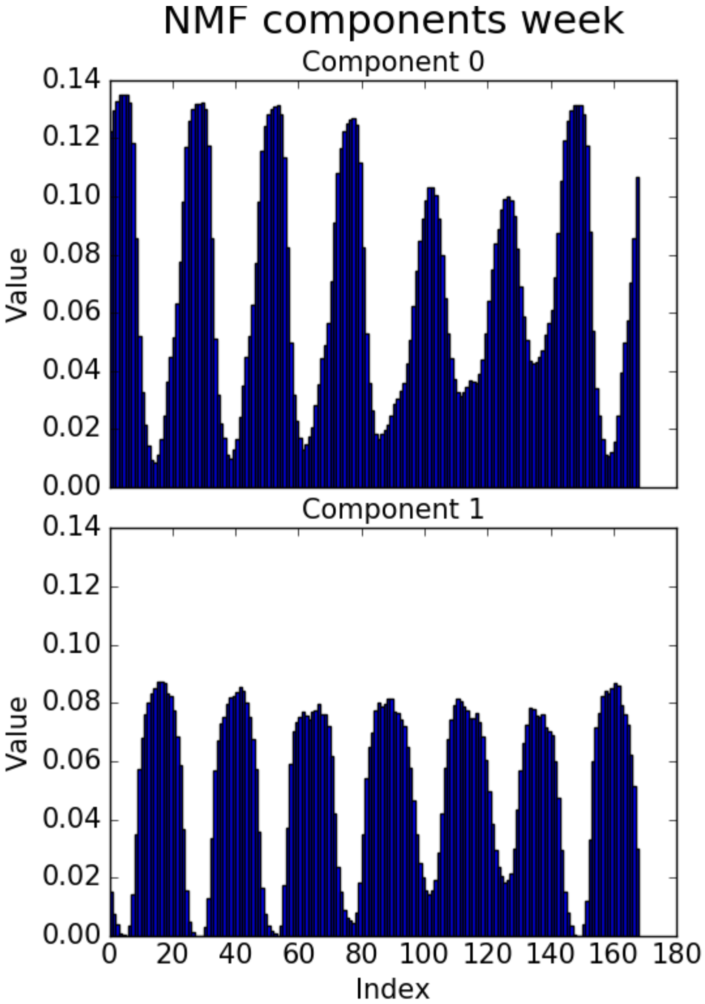

Figure 4.15: *The components of **H** for the weekly representation of **V**. Component 0 highlights the users that are mostly at home during the night hours of each day and Component 1 highlights the users that have the opposite behaviour, namely those that are the most at home during the daily hours on each day. These components resemble the components given for a daily resolution seen in Figure 4.12.*

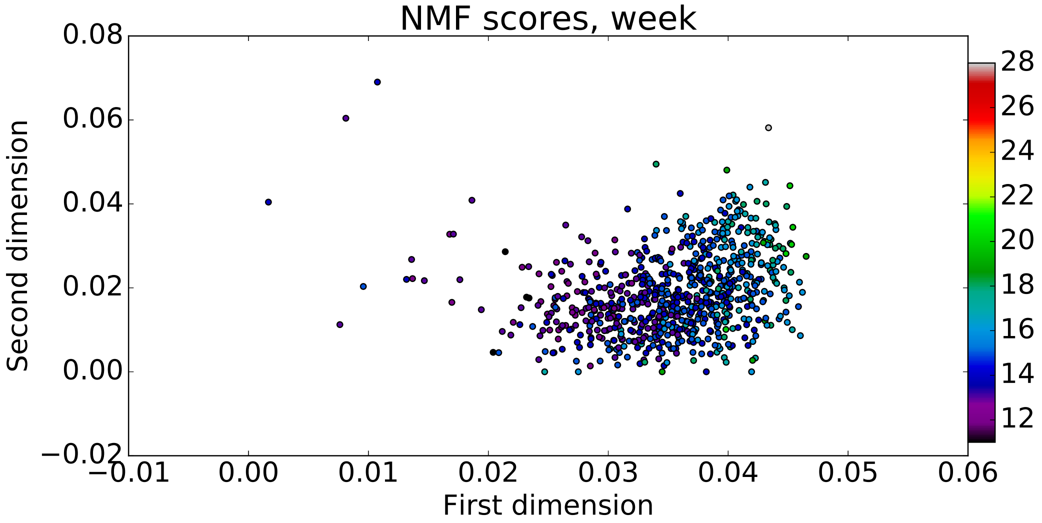

Figure 4.16: *The scores of the weekly NMF. The colour of the points correspond to the 95th percentile found from Scenario 3. As for the daily resolution given in Figure 4.13, a general trend is given that the higher the value in the first dimension, the larger the 95th percentile, but not a large deviation from the daily resolution is given.*

4.15 the components given in **H** can be seen. From Figure 4.15 it can be seen that there is a pattern that highlights the users' behaviour. It can be seen that in Component 0 the users that are mostly at home during the hours of the nights on a weekly basis are represented, and Component 1 represents the users that are mostly at home during the daily hours of the week. These results are closely related to the NMF results seen for the daily representation of **V**, however here the underlying structure on a weekly basis can be seen, which is also interesting. Not surprisingly however, the patterns that have been found are for users that are either mostly at home during the night or during the day, which also hints that no more components are needed, since there may not be other special types of users.

Again the scores found in **W** can be plotted to see if the weekly representation has the same properties as the daily representation in terms of findability. Such a figure can be seen in Figure 4.16. Again, the 95th percentiles for Scenario 3 has been used. Looking at Figure 4.16 it can be seen, similar to the daily analysis, that generally the higher the value in the first dimension, the higher the value of the 95th percentile. This is also the case for the daily representation, and also in this case it makes sense, since the first dimension represents common behaviour, namely staying at home during the night time of the weekdays. Therefore, the more common a user is, the harder they are to detect, and that is why the value of the 95th percentile is raised.

The groupings can now be inspected again to see the change from daily to weekly binning. A figure showing the colouring of the two groups in NMF space can be seen in Figure 4.17. Looking at Figure 4.17 it seems as though the representation on a weekly basis is quite similar to the daily representation, and may not yield any more information. This is of course only a qualitative measure, and should be verified using a classification method. Both the daily and weekly representation of the NMF will be used for classification to see which perform better, which can be seen in Section 4.3.

In this section the results obtained from the NMF method was given. The results were based on daily and weekly resolutions of **V**, which both tended to divide users in groups, where the first group are users who are the most at home during the night, which is the general trend, and users who are most at home

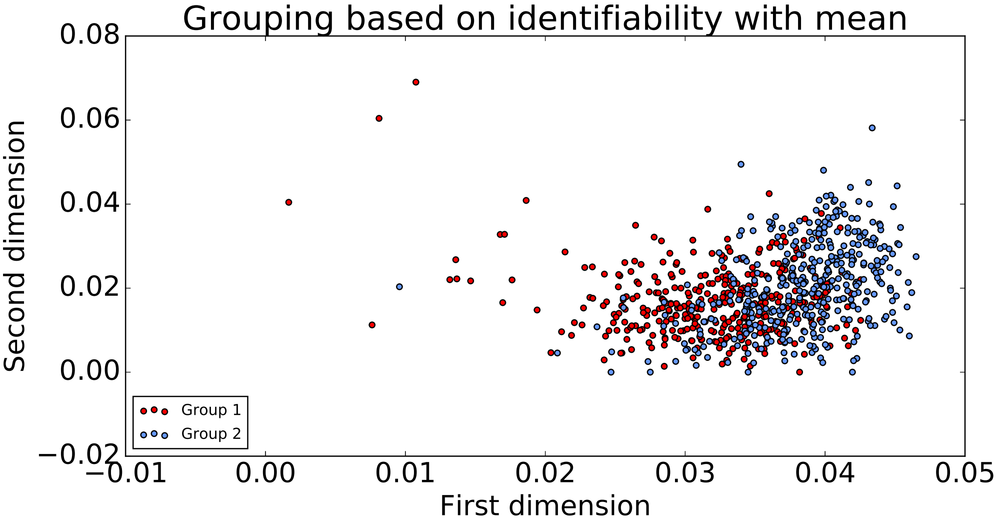

Figure 4.17: *Group 1 and group 2 coloured in NMF space for the weekly representation. The plot is very similar to the plot seen in Figure 4.14 for the daily resolution, but there are small changes when inspected closely. It seems though that the users may be classified based on their NMF scores.*

during the daily hours. Results were not that different, however it seemed as though users who had a larger value in the first dimension for both resolutions also had a higher 95th percentile, which can be seen in Figures 4.13 and 4.16. The difference of the resolutions will be quantified in Section 4.3 where users are classified based on their NMF scores in the daily and weekly resolution, respectively.

4.3 Classification of groups

In this section a classification of the groups defined by users' findability will be given. The classification will give insight in how much information the NMF results have, which are given in Section 4.2, when comparing to users' findability. The groups that are being classified are the groups found in Section 4.2 and the classification methods used are described in Section 3.3.

To start with, a K-Nearest Neighbour classifier has been used on the daily scores obtained from the NMF. To find the optimal number of neighbours, cross-validation has been used. To make things simple, a 10-fold cross-validation has been performed on a different number of neighbours, and the mean scores have been calculated. The scores are based on the percentage of correctly classified users for each fold, and then the mean has been taken for all folds. Cross-validation has been performed on 5, 10, 20, 30 and 50 neighbours, which yield the scores seen in Table 4.1. As a result, 30 neighbours has been used in the

Table 4.1: *Cross-validation scores for K-Nearest Neighbour classifiers for the daily representation. Here 30 neighbours yields the largest mean score.*

Number of neighbours	Mean score
5	0.738
10	0.719
20	0.731
30	0.740
50	0.737

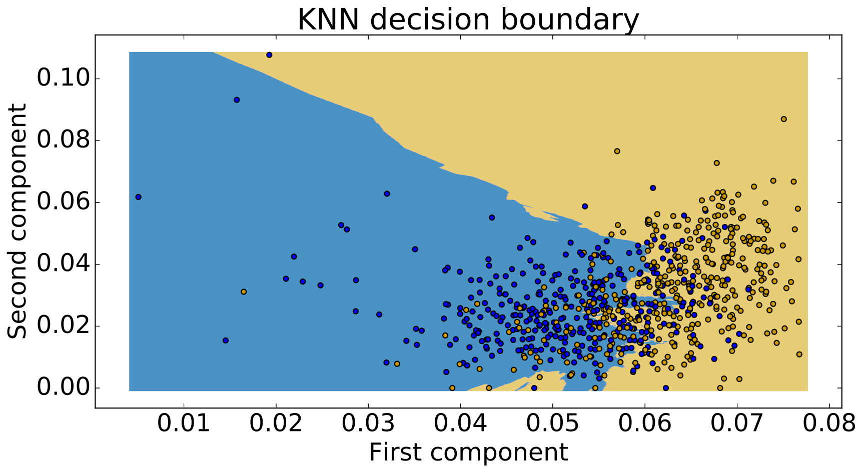

Figure 4.18: *Figure showing the decision boundary of the KNN classifier when choosing 30 neighbours for the daily representation.*

Table 4.2: *Cross-validation scores for Random Forest classifiers for the daily representation. Here 10 estimators yields the best mean score.*

Number of estimators	Mean score
10	0.719
100	0.705
200	0.703
500	0.707
1000	0.705

classifier. To get a sense of how the classifier works, one can see the decision boundaries of the classifier in Figure 4.18, which hold the same classes as Figure 4.14. Looking at Figure 4.18 one can see that the classes are not completely separated, but this is not possible without overfitting. If overfitting has been made, each point would have it's own little area around it holding the same color as it's label, but with the optimal number of neighbours chosen through cross-validation, the results will look as Figure 4.18. To find out how good the classifier performs on unknown data, a training and test set has been found, with a test set size of 10 % of the original data. Using the training set to train the classifier with 30 nearest neighbours and the test set to test the performance, a score of 0.789 is found, which definitely is acceptable. This means that in approx. 79 % of the time right label of a user is found, based on their NMF scores.

Another classification approach is the Random Forest classifier, which is an ensemble method, meaning it combines many classifiers (decision trees) and makes a new classifier based on these. Again, 10-fold cross-validation has been used to find the optimal number of classifiers/estimators. In Table 4.2 one can see the scores for a different amount of estimators in the Random Forest classifier. From Table 4.2 one can see that in this case, 10 estimators yields the best mean score. Therefore, a classifier with this number of estimators will be used to estimate the labels of users. When using the same training and test set as for the KNN classifier, a score of 0.711 is given for the Random Forest classifier. This means the KNN classifier is better than the Random Forest classifier, which may be caused by the dataset being better separable by KNN than Random Forest.

The weekly representation of the NMF will now be used to classify users with the same classification

Table 4.3: *Cross-validation scores for K-Nearest Neighbour classifiers for the weekly representation. Here 30 neighbours yields the best mean score.*

Number of neighbours	Mean score
5	0.733
10	0.736
20	0.744
30	0.751
50	0.744

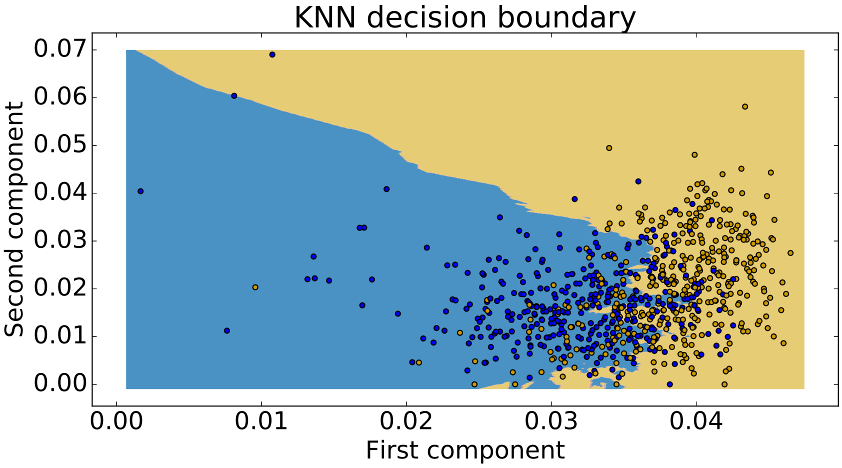

Figure 4.19: *Figure showing the decision boundary of the KNN classifier when choosing 30 neighbours for the weekly representation.*

methods used as for the daily representation. For the KNN, the optimal number of neighbours is determined from Table 4.3. From Table 4.3, 30 neighbours is chosen as the optimal choice for this parameter. The decision boundary of the classifier can be seen in Figure 4.19. Figure 4.19 resembles Figure 4.18, but with some alterations. Again, it is worth noting that the number of neighbours chosen does not cause overfitting when inspecting the figure.

A training and test set of the same size as for the daily representation will also be used to find the performance of the KNN for the weekly representation. With the given training and test set, a score of 0.816 is found, which is a good result. Comparing this result to the daily representation, the classifier performs better with the weekly representation than the daily, which is interesting.
A Random Forest classifier has also been used for the weekly representation, and the number of estimators is based on the results in Table 4.4. From Table 4.4 the optimal number of estimators is chosen to be 1000. Given this parameter choice and the training and test set from the KNN classifier, the performance is found to be 0.763. This is not as good as the KNN classifier, which was the same case as for the daily representation. However, this classifier is still better than the Random Forest classifier for the daily representation, so in general it seems that a weekly representation works better when classifying the groups.

From the results seen in this section, it seems plausible that NMF is a good separator of the classes, which are defined by the 95th percentile. Since the scores of the classifiers range from 0.711 as the worst to 0.816 being the best, it is clear that there is some correlation between the 95th percentile and the NMF representations. As a general consideration, there are a lot of parameters to tune in the Random Forest classifier, so these may have an influence on the performance as well. The main point however is

Table 4.4: *Cross-validation scores for Random Forest classifiers for the weekly representation. Here 1000 estimators yields the best mean score.*

Number of estimators	Mean score
10	0.739
100	0.734
200	0.740
500	0.742
1000	0.743

that from two basic classifiers the obtained scores are quite good, without being prone to overfitting due to the use of cross-validation.

4.4 Likelihood

In this section the results obtained from the likelihood analysis will be presented. The method for finding the likelihood is presented in Section 3.4.
The likelihoods are based on the distributions of the probability of observing a 1 for all users (or the fractions of 1's for the available data for all users). This means that all observations are binned in either a resolution of 24 or 168, meaning 24 bins for a daily basis or 168 bins for a weekly basis, respectively. Then, for each bin the probability of observing a 1 is found for all users.

To get an impression of what the distributions look like, the median together with the lower and upper quartile has been plotted for all of the distributions. For the daily resolution, one can see such a plot in Figure 4.20. It should be mentioned that users with less than 10 % valid data have been removed, since keeping them does not make much sense for the likelihood analysis. Looking at Figure 4.20, one can see

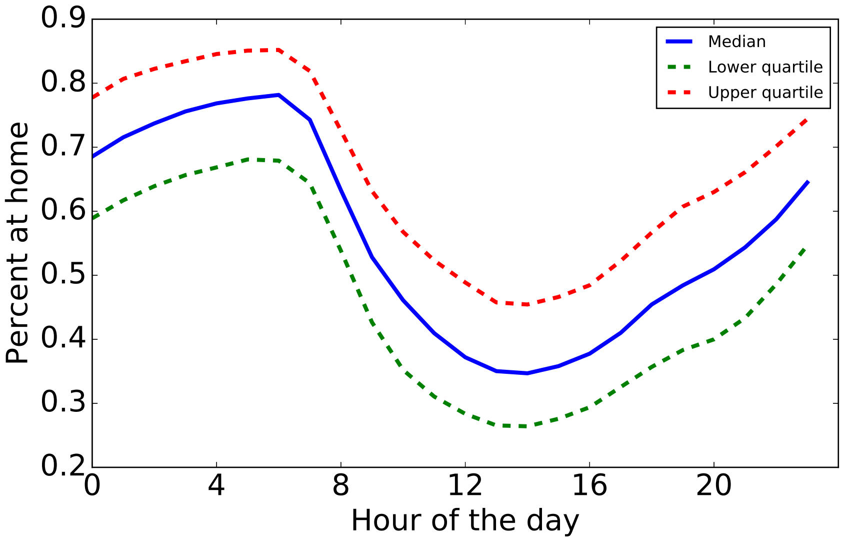

Figure 4.20: *Median, lower and upper quartile for each distribution of percentages of 1's with a daily resolution. It can be seen that users deviate more in the middle of the day where the distributions are the widest than any other given time. Users also deviate the least during the morning hours where the distributions are the narrowest.*

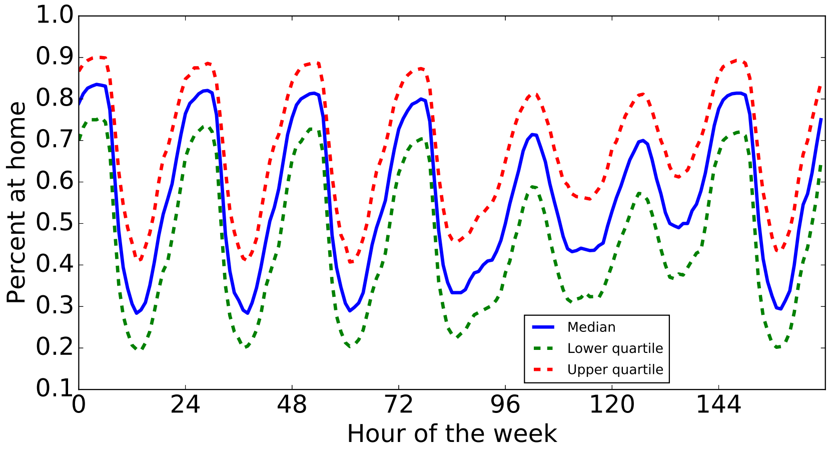

Figure 4.21: *Median, lower and upper quartile for each distribution of percentages of 1's with a weekly resolution. It can be seen that each morning users deviate the least and in the middle of each day users deviate the most. A special case occurs on Fridays and Saturdays where users deviate more in general than any other day because of parties and other social gatherings.*

that the typical behaviour on a daily basis is being at home during the night and away from home during the day. It can also be seen that the broadest distribution is found in the middle of the day, while the narrowest distribution is found in the morning. Therefore, users deviate more in their behaviour during the middle of the day and the least in the morning.
In Figure 4.21 one can see the distributions for a weekly resolution. From Figure 4.21, it can be seen that there is a daily pattern for users' mobility. It is also clear that generally the distributions are the narrowest during the morning of each day, and widest during the middle of the day. However there is a special case during the fifth and sixth days, that is Friday and Saturday. Here the distributions are wider, meaning that users are having very different behaviour. Users being students, this means that some students are attending parties or social gatherings during these days, while others are not.

The likelihood can now be found for both the daily and weekly resolution. In Figure 4.22 one can see the log likelihoods for all users for the daily and weekly resolution. Looking at Figure 4.22, the log likelihoods are very negative, meaning very small likelihoods, which may be caused by the normalization of the histograms. Furthermore, the daily and weekly log likelihoods seem to be highly correlated. To verify this, a plot has been produced that shows the log likelihood for the daily resolution on the first axis and for the weekly resolution on the second axis. Such a plot can be seen in Figure 4.23. From Figure 4.23, it is clear that the two likelihood estimations are highly correlated, which means that both can be used for the analysis. The *Spearman correlation* has been computed, which compared to the Pearson correlation does not assume a normal distribution of either dataset. The function returns a correlation of 0.909 with a p-value of $1.087e-288 \approx 0$ with a null hypothesis that the datasets are uncorrelated [31]. From these results it can be argued that either the daily likelihood or the weekly likelihood can be used, since they are highly correlated, of which the daily likelihood estimates will be used further on.

It would be expected that the likelihood is correlated with the findabilitiy. Intuitively, the harder it is to find a user, the larger their likelihood. This is because if a user is typical, they require a large number of questions to be asked if they should be found (high 95th percentile), as well as having a large likelihood value, since they will be likely to come from the distributions. In Figure 4.24 one can see a plot showing the log likelihood versus the findability of users. In this case the 95th percentiles for Scenario 3 has been chosen, as this is the distribution of 95th percentiles with the least variation, as seen in Figure

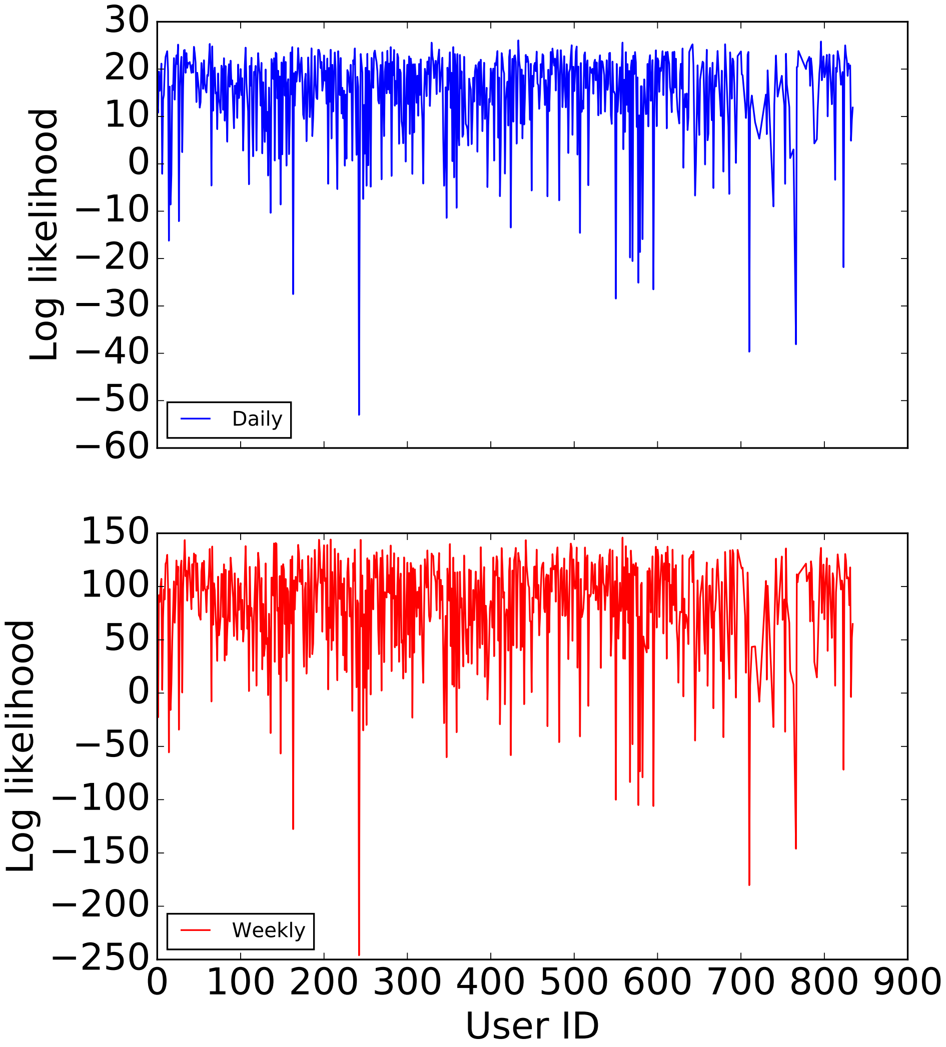

Figure 4.22: *Log likelihoods for all users for the daily and weekly resolutions. The likelihoods in general have large negative values, which may be caused by the normalization. It can also be seen that the likelihoods seem to be very correlated given the different resolutions.*

4.7. Inspecting Figure 4.24 it seems as though there is some correlation, however not much. Finding the Spearman correlation between the log likelihood and the findability yields a result of -0.124 with a p-value of $6.34e - 4$. This means that if the log likelihood is increased, the 95th percentile is slightly decreased. This seems rather odd, since it would be more probable that the higher the 95th percentile or the lower the findability, the more common a user is.

From the likelihood analysis shown in this section, the results indicate that when finding the likelihoods for the users, the resolution does not matter since they are highly correlated. The results also indicate that the higher a user's 95th percentile, the lower the log likelihood, as seen in Figure 4.24, which is not in correspondence with the intuition. The results may therefore suggest that this method of likelihood estimation is not applicable in this context or that the construction of the distributions need to be altered. The results will be discussed in detail in Section 8.1.

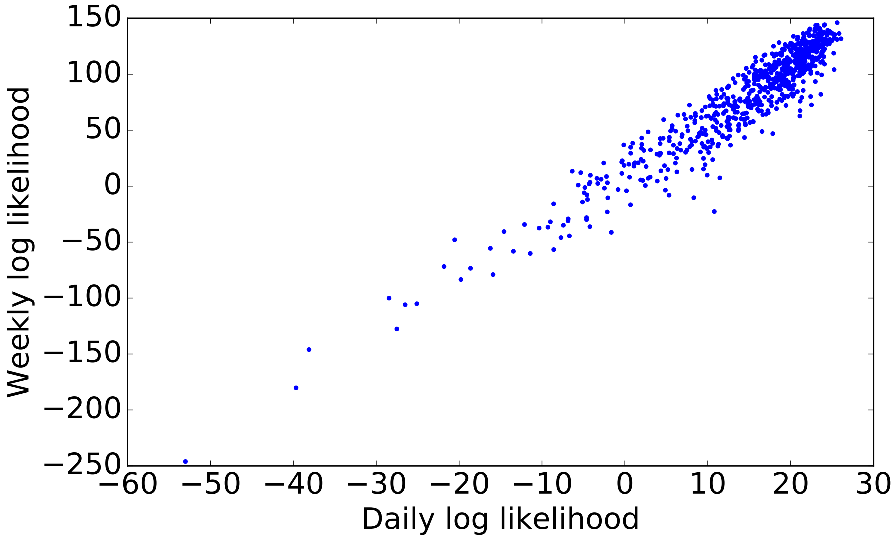

Figure 4.23: *Plot showing the correlation between likelihood in a daily and weekly resolution. The log likelihoods are highly correlated with a Spearman correlation of 0.909 and shows that either of the likelihood distributions can be used.*

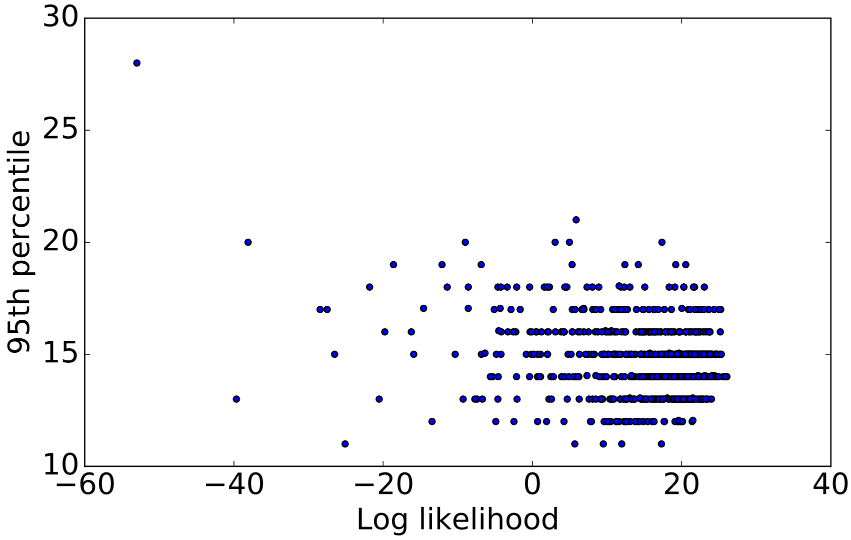

Figure 4.24: *Plot showing the correlation between log likelihood and the 95th percentile for all users. Here the 95th percentile has been found from Scenario 3. The correlation between the daily log likelihood and the 95th percentile is found using the Spearman correlation to be* -0.124*, meaning there is a small negative correlation. This however is in contradiction to the intuitive belief that if a user is more likely to come from a distribution, they should be harder to find, but the results state the opposite.*

4.5 Temporal correlations

In this section the results from the temporal correlation analysis will be discussed. The results will indicate which time bins are the most informative when identifying users. The analysis will be performed in a daily and weekly resolution, respectively. The results are found by applying the method described in Section 3.5.

For all users, the number of users who are not equal to the index user divided by the number of users not equal to the index user at the previous time bin has been found. The fraction is 1 if no users were discarded in the time bin and 0 if all users were discarded in the time bin. Here the mean has been taken in all time bins and for Scenarios 1 and 3, users with less than 10 % valid data have been removed, where users with less than 50 % have been removed for Scenario 2. This is done such that all users can be found, which would not be possible otherwise. The results for all scenarios for a daily and weekly resolution can be seen in Figure 4.25. From Figure 4.25 it is clear to see that some time bins hold more information than others. For example, looking at the results for the daily resolution in the left figure, it seems better to discard users using time bins that are in the night and early morning rather than in the afternoon. For the weekly resolution in the right figure, it also seems better to use information from the night time rather than the afternoon. However, during Friday and Saturday it can be hard to discard

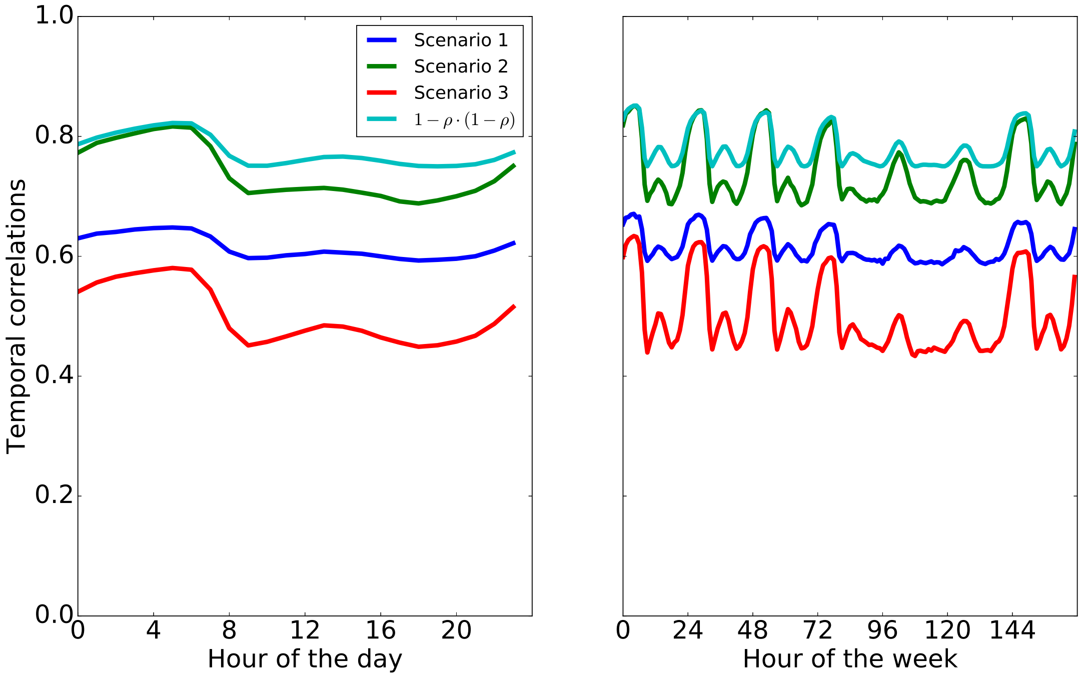

Figure 4.25: ***Left figure****: Plot showing the temporal correlations on a daily basis together with the daily density curve. It can be seen that the time bins that hold the most information when discarding users are the ones in the night time for all scenarios.* ***Right figure****: Plot showing the temporal correlations on a weekly basis together with the weekly density curve. As for the daily resolution, the bins that hold the most information are the ones in the night time for all scenarios. A special case occurs for Friday and Saturday however where not much information is carried with the time bins. In both figures, all users have been included and the mean has been taken. Additionally, the curves for each scenario in each resolution are following the density curve quite closely, only differing by a proportional factor as expected. This means that when averaging over all users, the temporal correlations do not show any significant difference between the three scenarios, which means that users may not be treated differently.*

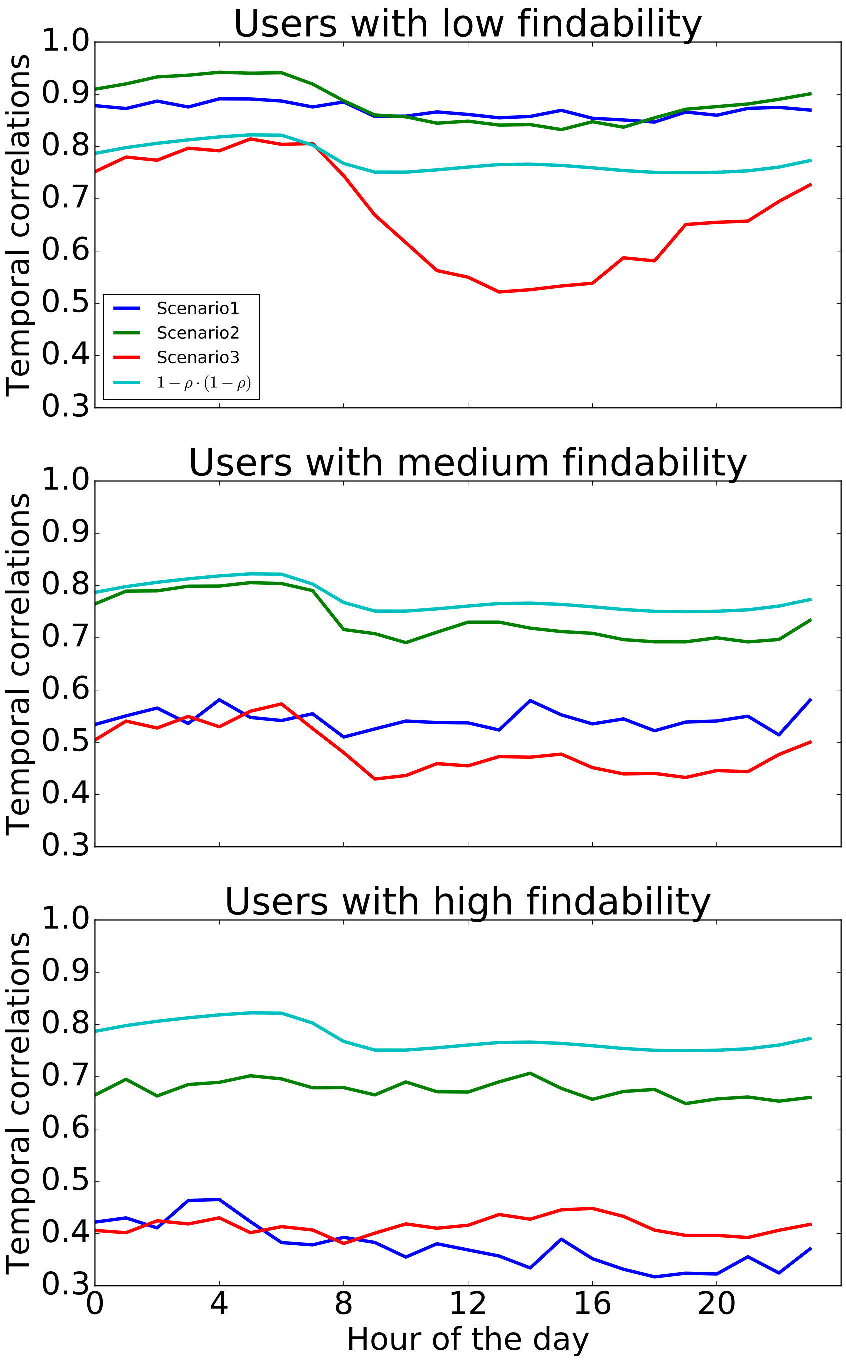

Figure 4.26: *Plot showing the temporal correlations for users with low, medium and high findability with a daily resolution. A higher findability generally yields less information in the time bins, which is evident when looking at the curves for all scenarios, since raising the findability yields a more flat curve. Raising the findability also lowers the curves, which makes sense since a low value on the second axis means more users are discarded at each iteration.*

users no matter which time bin is being used during these days. It is also clear if Scenario 2 is chosen, it is harder to identify users, where Scenario 1 is easier and in turn easiest using Scenario 3. These results are in correspondence with the unicity results seen in Section 4.1.

The comparison between the density curve $1 - \rho \cdot (1 - \rho)$, which is described in Section 3.5, and the temporal correlations can also be seen in Figure 4.25. Looking at Figure 4.25, it can be seen that the functions follow each other closely and are only different to a proportional factor, depending on the scenario chosen, which is also what is expected. These results yield that when averaging over all users, a significant difference cannot be found given the different scenarios. It can only be seen that the different scenarios give different results, but it seems that users are not treated differently as they indeed should be.

To investigate further, users which have high, medium and low findability will be analysed in the same way. This is done such that the difference of users will be highlighted. To brush up, high findability means a low 95th percentile and low findability means a high 95th percentile. As the scenarios are quite different, the users with high, medium and low findability, respectively, will have different values of the 95th percentile. Also, the findabilities will be distributed in different ways, since for example Scenario 2 yields hard identifications of users. This analysis is therefore performed to highlight the individual behaviour of users and their temporal correlations.

For all scenarios, the users with the lowest, median and highest findability has been found and the temporal correlations have been found. The users that have been found are not the same, so different users are analysed for the different scenarios. It should be mentioned that of all the fractions of discarded users in a given time bin (daily or weekly resolution) the mean of those fractions has been used in the analysis. The results for a daily resolution for individual users, can be seen in Figure 4.26 together with the density curve. From Figure 4.26 it can be seen that for all types of users, the individual temporal correlations of Scenario 1 do not really change in shape. It can be seen that it is almost constant for any given user, however the points on the curve have a lower value when moving from a user with low findability to a user with high findability. For Scenario 2, the temporal curve almost resembles the density curve, however it is above the curve for a user with low findability and below for a user with medium findability. Finally, for Scenario 3 it is clear that some time bins are more important than others, especially for a user with low findability. For a user with high findability however, the curve is almost flat, so no further information is given from a specific time of day. However, it can be seen that for a user with medium findability the curve is not quite flat and therefore some importance of time bins is still present.

The results of the temporal correlation analysis for individual users for a weekly resolution can be seen in Figure 4.27 together with the density curve. The results given in Figure 4.27 show that users with low findability hold much more temporal information than others, since there are strong peaks in all curves and especially for Scenario 3. Also, it is easier to see that on a weekly basis, users with a low findability hold much more temporal information than was given for the daily basis seen in Figure 4.26. For users with a medium findability, some trends can be seen, however they are not as clear as for users with a low findability. The trends still follow the density curve somehow, especially for Scenario 2, however not perfectly. For users with high findability, the curves are almost constant, or at least there is no obvious trend that resembles the density curve. In general, the points on the curves have lower values the higher the findability, which makes sense since a low value means many users are discarded. Therefore, a high findability means more users are discarded in all time bins than for a low findability, which makes sense.

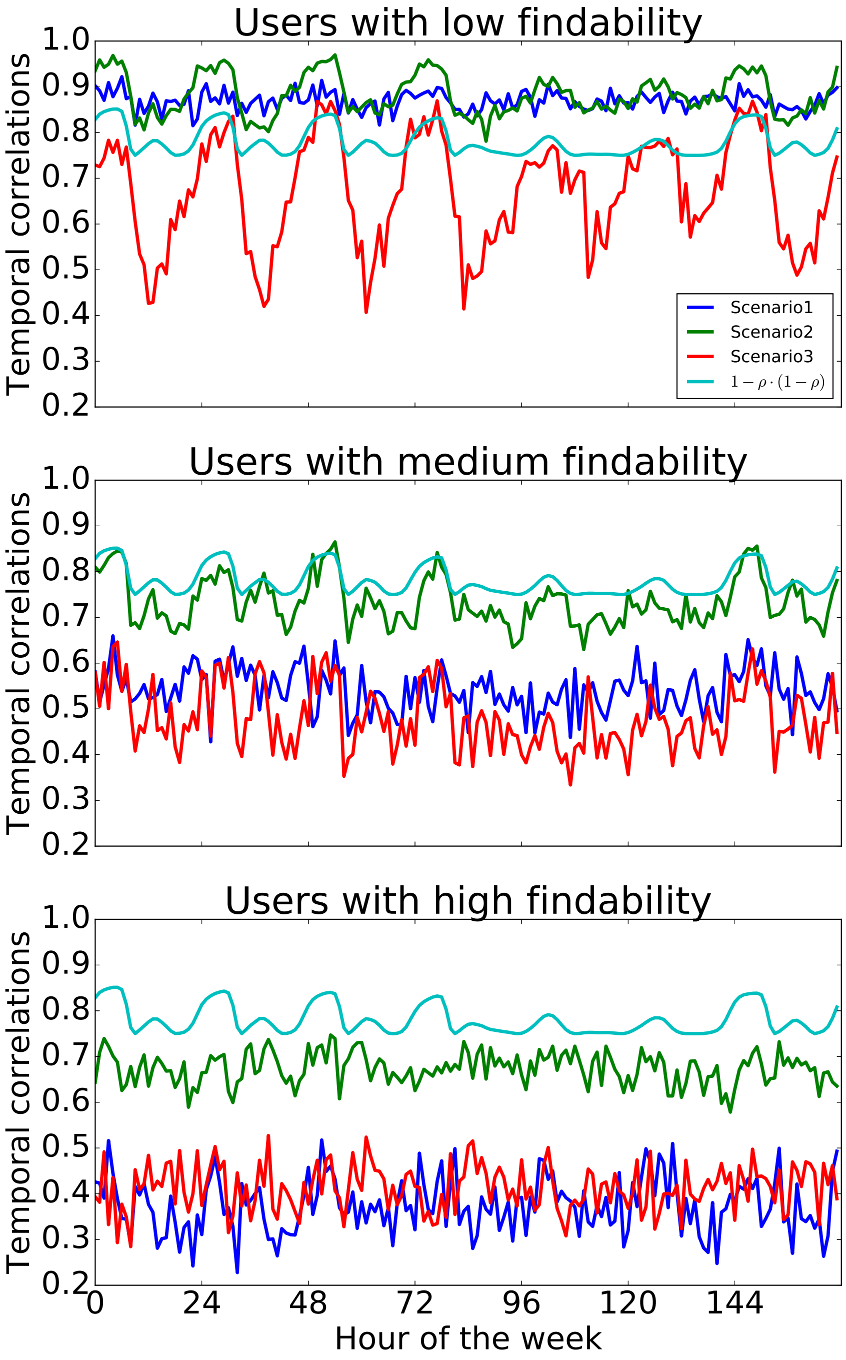

Figure 4.27: *Plot showing the temporal correlations for users with low, medium and high findability with a weekly resolution. Less information is given in the time bins when raising the findability, which also was the case for the daily resolution. Raising the findability also lowers the curves, as also was the case the daily resolution.*

To conclude, the results in this section show that indeed certain hours of the day or the week yield easier identification of users. From Figure 4.25, it is given that the nightly hours yield easier identification of users and a special case is given on Fridays and Saturdays which in general do not yield any information. Furthermore, looking at different users, users with a low findability are prone to having time bins that are more important than others, which especially is clear in Figure 4.27. If taking users with medium findability, the temporal information gets less important, and almost not at all for users with high findability. This is interesting, since users who are hard to find (low findability) can be found by looking at specific times of the day or week, but for users who are easy to find (high findability), it does not really matter which time bin the information comes from.

In this chapter analyses performed on the **Home/Work** dataset have been discussed. These included unicity simulations in Section 4.1, NMF in Section 4.2 with classification methods regarding the NMF features in Section 4.3, likelihood analysis in Section 4.4 and temporal correlation analysis in this section. Comparison to these results will be made in the other datasets in the following chapters.

Chapter 5

Social labels

In this chapter an analysis on the social datasets described in Section 2.3 will be discussed. The results from finding unicity will be shown and analysed in Section 5.2, as well as the temporal correlations with a daily and weekly resolution shown in Section 5.3.

5.1 Adjustment of social labels dataset

From early analysis, results indicate that when finding unicity for each social dataset described in Section 2.3, that is Facebook, call and sms datasets, it is very hard to identify users. A large number of points is needed to identify users for all datasets, which also can be seen in Figure 5.1 where the 95th percentile for all users for the sms dataset is given (see Defintion 1). It should be noted that the dataset has been processed such that the time stamps used are within the same range as the mobility dataset in Section 2.2 and users with less than 10 observations have been removed. From Figure 5.1 it is clear that few users have been identified, and those who are need a lot of points before they can be identified properly. As the results seen in Figure 5.1 are the best results given the three social datasets, better results cannot be found when looking at them separately. This may be caused by the datasets having few social observations for all users, where the sms dataset has the most social observations and the best results. Therefore, users may be close to identical, since only a few time bins stand out. The percentage of being social in each dataset can be seen in Figure 5.2, that is the number of 1's divided by the total number of time stamps for each user.

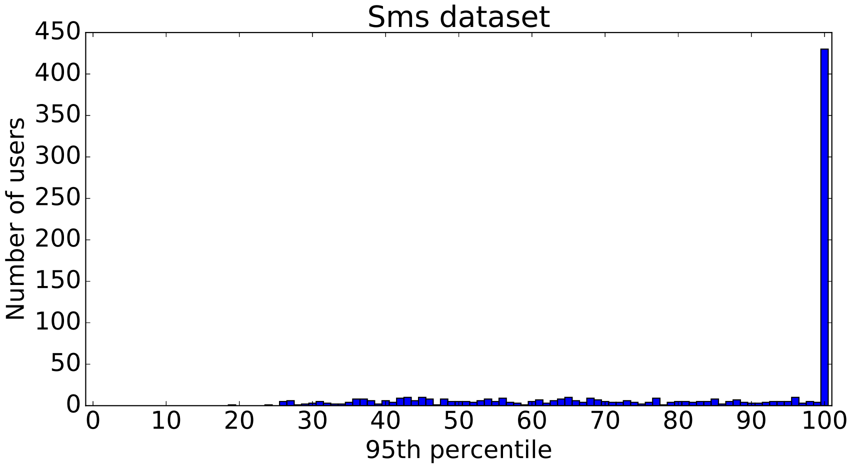

Figure 5.1: *Figure showing the 95th percentiles for the sms dataset. From the figure it is evident that users are hard to identify, and the ones who are identified need a lot of points to be found. A value of 100 means that the user has a 95th percentile of 100 or more, where the algorithm stops at 100 points if the user has not been found.*

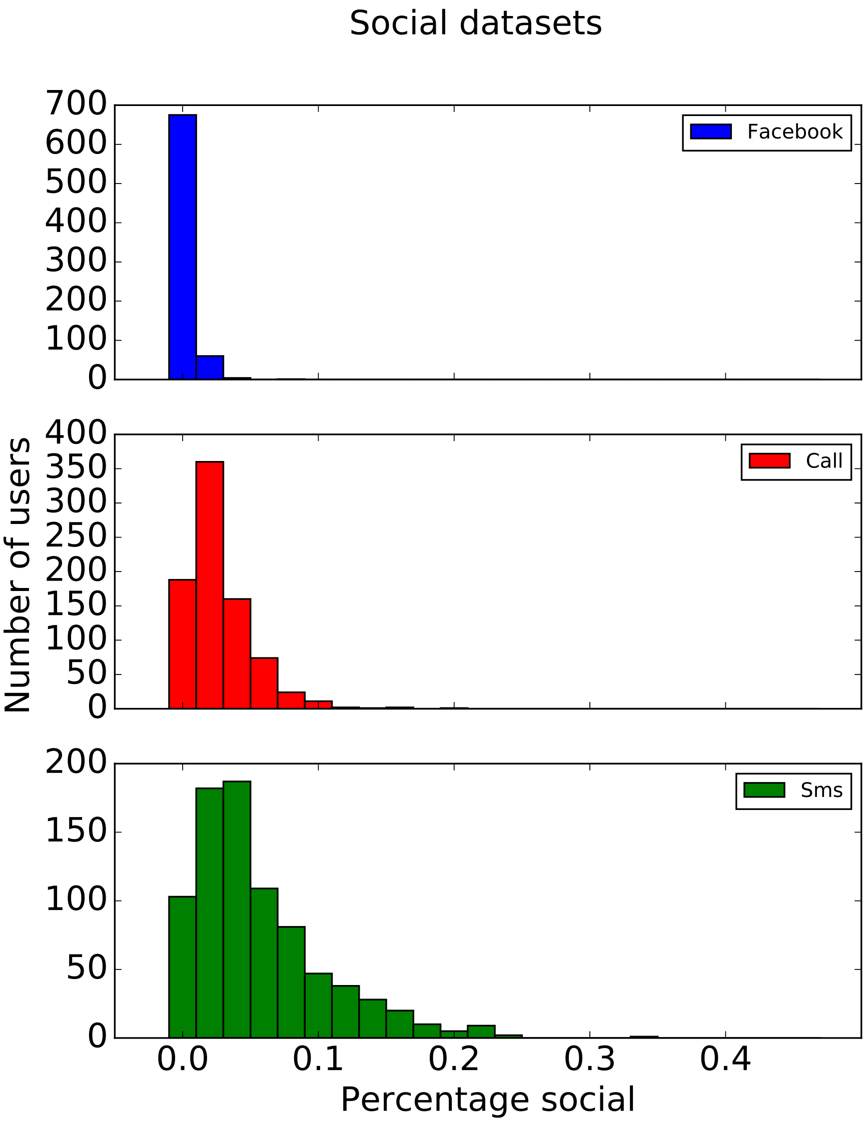

Figure 5.2: *Figure showing the percentage of all users being social. The percentage is the number of 1's divided by the total number of time stamps, which is different from each dataset. From the figure it can be seen that not many observations are available, yielding it hard to identify users. It can also be seen that the Facebook dataset has the least number of observations, the call dataset has more and finally the sms dataset has the most.*

From Figure 5.2 it can be seen that the sms dataset has the most observations, the call dataset has the second-most observations and the Facebook dataset has the least. Also, users do not in general have a large amount of observations, either from not using Facebook/call/sms that much or maybe from missing observations.

A workaround has been made, which takes the three datasets and combines them into a *single social dataset*. This has been done by taking all times bins that are shared in all datasets, as well as taking all users that are present in all datasets, and let these users be social if any Facebook/call/sms event is given, and not social otherwise. Social is given as a 1 and not social is given as a 0. Some users in the new

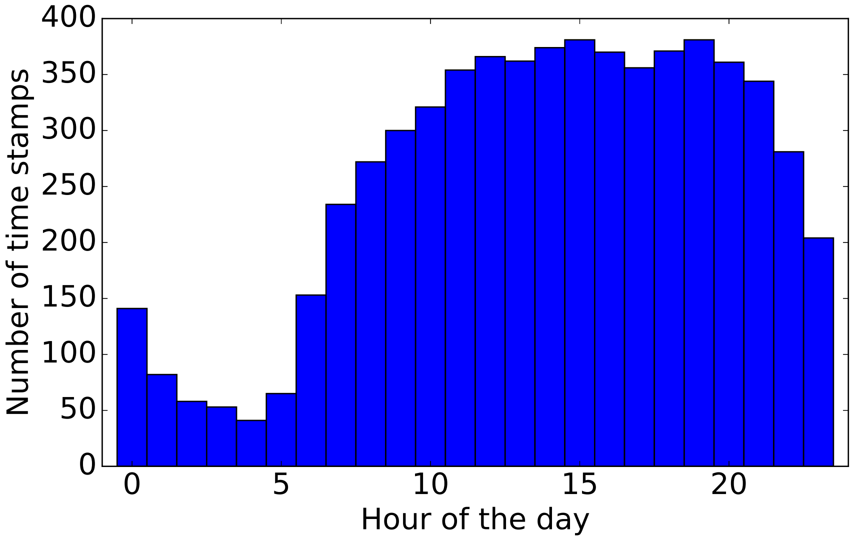

Figure 5.3: *Figure showing the distribution of time stamps for the new social dataset. It can be seen that there are a lot more time stamps in the late morning and afternoon hours with much less observations in the morning and night hours.*

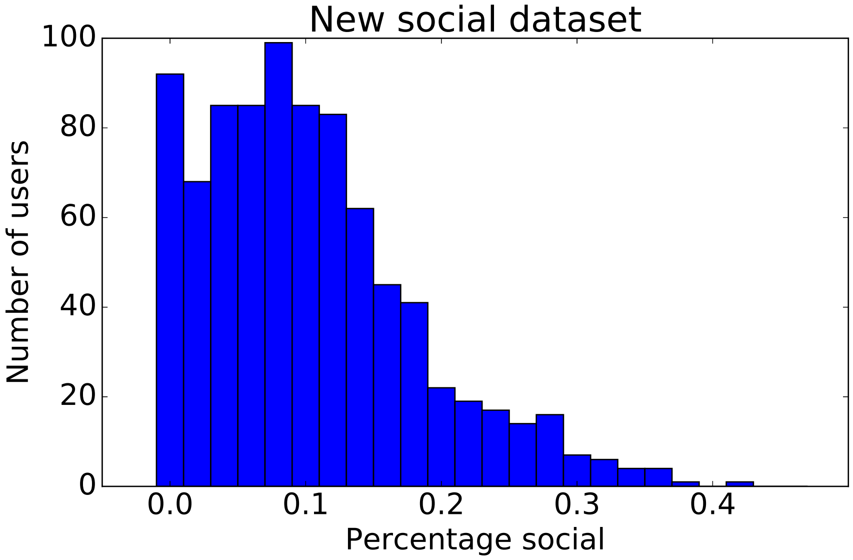

Figure 5.4: *Figure showing the percentage of all users being social in the new social dataset. The percentage is the number of 1's divided by the total number of time stamps. The figure indicates that the number of observations have increased for all users when combining the datasets, justifying the use of the combined dataset.*

dataset are duplicated, and removing them yields 856 users and 6225 time stamps. However, there is a skewness of the time stamps, since there are a lot more time stamps in the late morning and afternoon hours with much fewer in the early morning and night hours. This observation can be seen in Figure 5.3. The observation given in Figure 5.3 may have an influence on the identification of users, which will be discussed in Section 5.3.

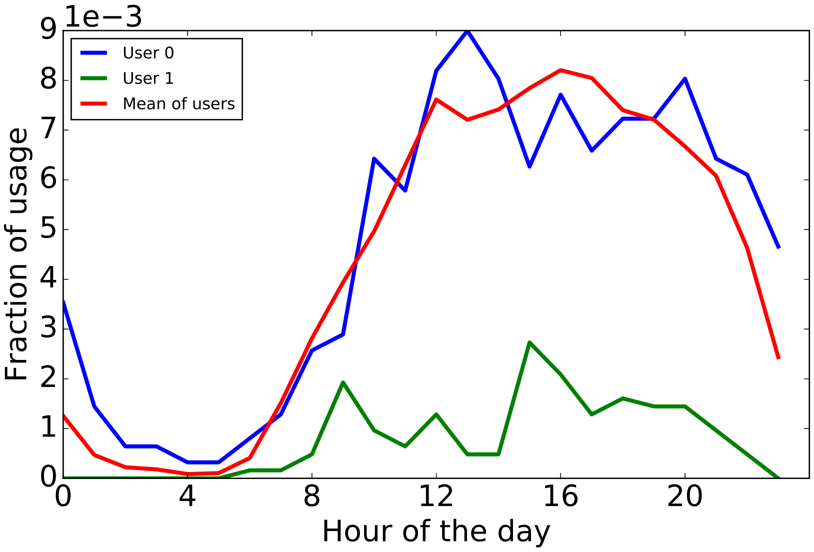

Figure 5.5: *Figure showing the fraction of usage for the* ***dialogue*** *dataset in a daily resolution. Here the usage is shown for the mean of all users, as well as individually for user 0 and 1. It can be seen that users are most socially active during the afternoon and early evening and the least in the morning, which corresponds to each individual dataset seen in Section 2.3. As indicated by user 0 and 1, users can also be quite different in their social activity.*

Figure 5.4 shows the percentage users are social in the new social dataset. By comparing Figures 5.2 and 5.4 it would seem that the fraction of observations in general have been increased for all users. It should be noted however that the number of time stamps and users all differ in the three social datasets, so it can be hard to compare the *old* datasets with the new, but the shape of the distribution in Figure 5.4 shows the improvement. Therefore, the new combined social dataset will be used in further analysis. It will from this point be addressed as the **dialogue** dataset.

In Figure 5.5 one can see the fraction of usage for the **dialogue** dataset in a daily resolution for two users and the mean of all users. Looking at Figure 5.5, one can see users are most socially active in the afternoon and the least active during the night, which corresponds to the behaviour of each individual dataset seen in Section 2.3. Furthermore, Figure 5.5 indicates that users can be quite different when comparing user 0 and 1. In Figure 5.6 the fraction of usage can be seen in a weekly resolution for the same users and the mean of all users. Looking at Figure 5.6 one can see that the behaviour is the same as for the daily resolution in general, namely that users are the most active in the afternoon hours and the least in the nightly hours. Again, comparing users 0 and 1 gives different kinds of social activity. User 0 has a pattern that resembles the mean user, with a bit more social activity on Fridays and Saturdays, where user 1 has the most social interactions on Fridays and not much activity the rest of the week.

In the following sections, the results from the unicity simulations and the temporal correlation analysis will be given on the **dialogue** dataset.

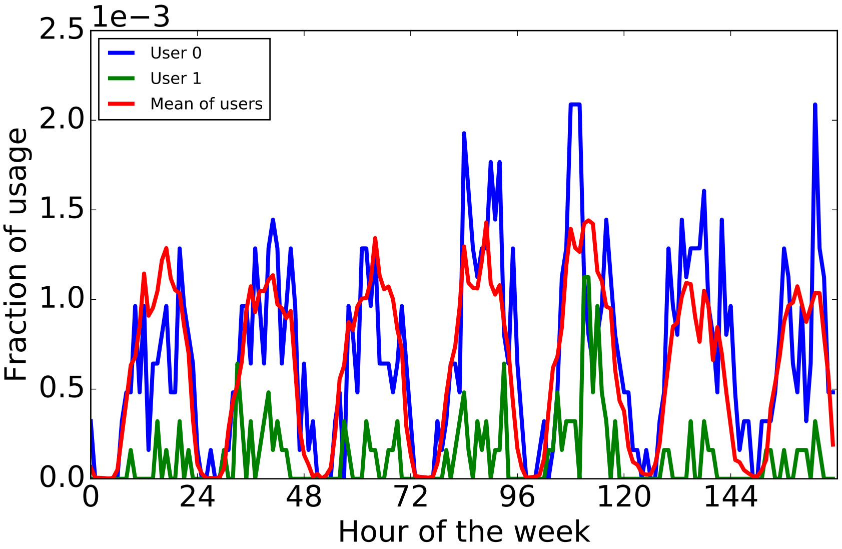

Figure 5.6: *Figure showing the fraction of usage for the* ***dialogue*** *dataset in a weekly resolution. Here the usage is shown for the mean of all users, as well as individually for user 0 and 1. Users have a specific pattern, where they are mostly active in the early afternoon and night hours, which is similar to the daily resolution seen in Figure 5.5. Furthermore, users 0 and 1, as also indicated in Figure 5.5, are quite different.*

5.2 Unicity

Using the attack model defined in Section 3.1, the 95th percentile for all users has been found. It should be mentioned however since no missing data explicitly exists, meaning no obvious placeholder values for missing data are present in the dataset, not all scenarios are discussed. This is because all scenarios

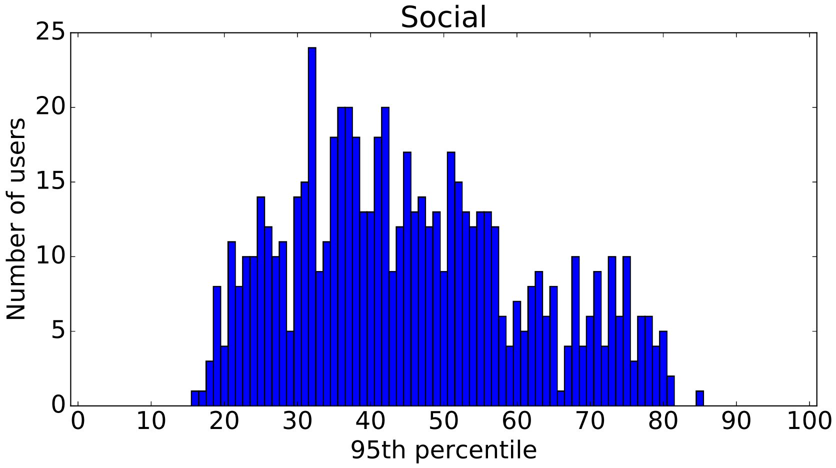

Figure 5.7: *Figure showing the results from the unicity simulation for the* ***dialogue*** *dataset. The figure shows a histogram for the 95th percentiles of all users who have at least 250 observations. Comparing the results with the results from the* ***Home/Work*** *dataset, it is given that users are harder to find using social labels. The distribution of the 95th percentiles is quite broad in comparison to the distributions found for the* ***Home/Work*** *dataset, indicating users can be quite different in their social behaviour.*

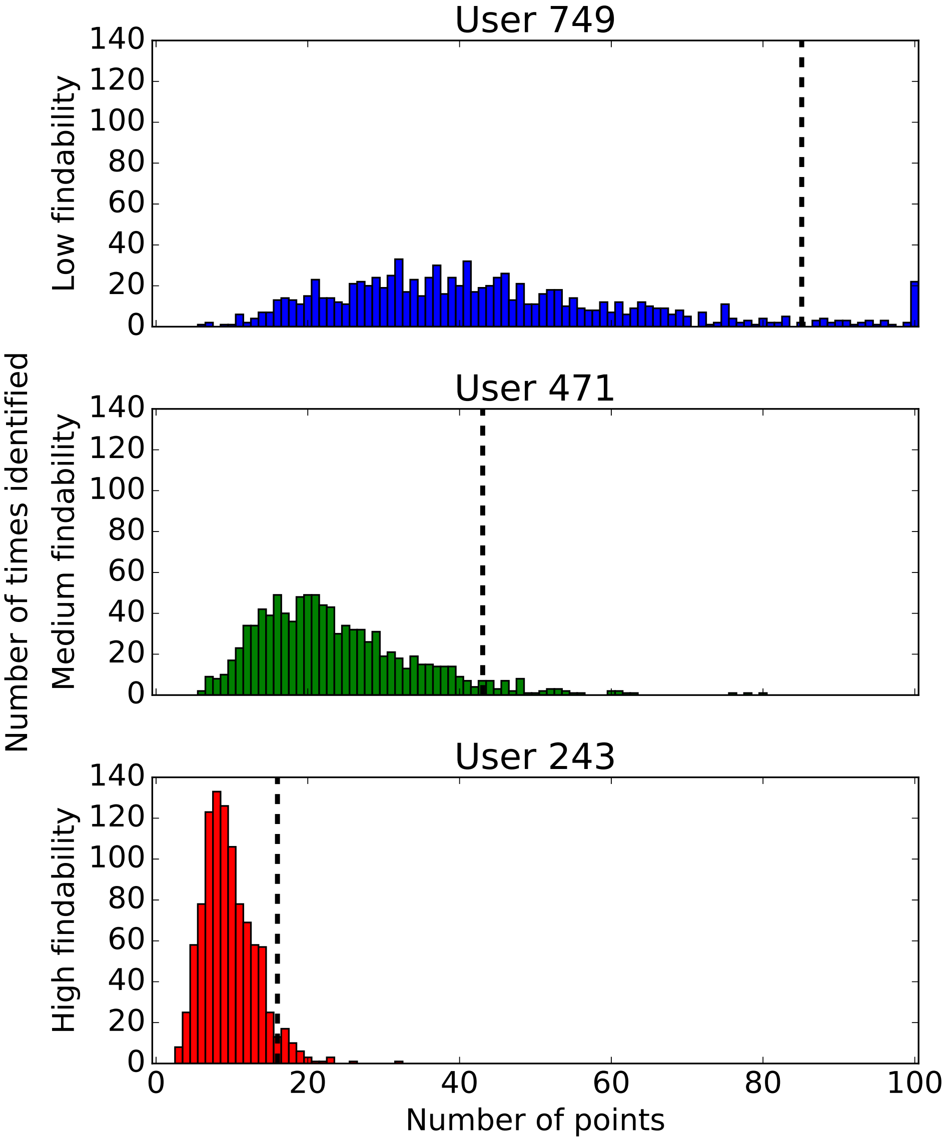

Figure 5.8: *Figure showing histograms for identifications for users with different findabilities. The black dashed lines show the 95th percentiles. It can be seen that user 749 has not been identified in all iterations, because of the observations at 100 points. It is also evident that the higher the findability, the more narrow the distribution is. Therefore, if a user is easy to find then in general there is not much variation in how many points are needed to identify the user. For users with low findability, there is a much larger spread in the number of points needed to identify a user.*

are equal if no missing data exists. To get decent results and not to have users who lack too much data, users who have less than 250 observations are removed. This preprocessing yields 659 users and 6225 time stamps. The results from the simulations can be seen in Figure 5.7. Comparing Figure 5.7 with the results for the **Home/Work** given in Section 4.1, it seems as though it is harder to identify users using social labels instead of mobility labels. The distribution seen in Figure 5.7 is also broader than the distributions found for the **Home/Work** dataset, indicating that some users are easy to find and some are hard to find. Therefore, users are very different in their social behaviour and not as much in their mobility behaviour. This may be due to the fact that people's mobility patterns are more regular than their social patterns. Users may not call, send text messages or use Facebook regularly on specific hours

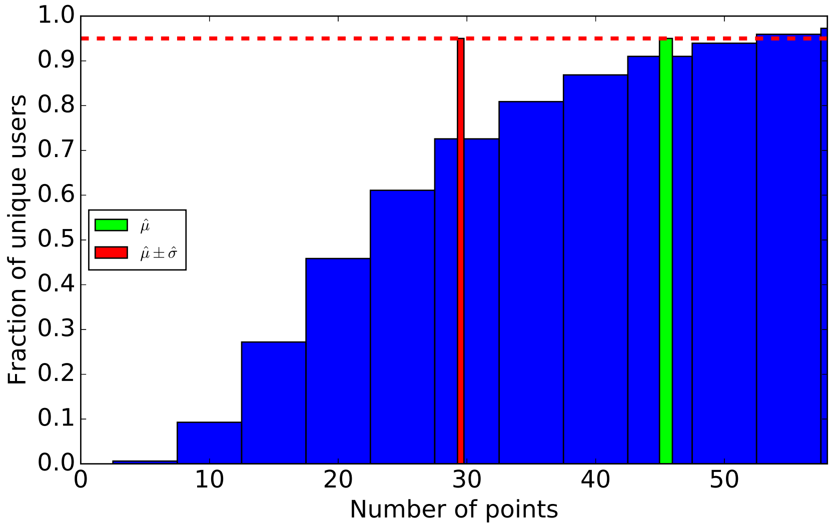

Figure 5.9: *Figure showing the mean fraction of users identified at each number of points for the three scenarios over all iterations. The green bar shows the mean 95th percentile and the red bars show the spread of the standard deviation from the mean. It can be seen that the mean of the 95th percentile almost equals the mean number of points to identify 95 % of users, however with deviating about 5 points and a large spread in the distribution.*

of the day on schedule, where users travel in a specific pattern from home to work on weekdays. It may therefore be argued that users are harder to identify based on their social activities, here Facebook, call and sms events, rather than their mobility patterns.

To get an idea of how each individual is being identified within the social context, a figure showing the spread of identifications for different users is seen in Figure 5.8. From Figure 5.8 it is given that the higher the findability, the lower the spread of points needed to identify a person. This can be seen by the shape of the distributions, as the distributions become more narrow when the findability is higher. This was also the case for the **Home/Work** labels, as seen in Figure 4.8. This in turn means that the higher the findability, the more a *typical* amount of points is needed to identify a user. Furthermore, user 749 has not been identified in all iterations, but still has a 95th percentile below 100 points.

A figure showing the mean fraction of users uniquely identified at different number of spatio-temporal points similar to the figures presented by de Montjoye et al. can be seen in Figure 5.9, which is similar to Figure 4.9 for the **Home/Work** labels [14]. As stated by de Montjoye et al., only four spatio-temporal points holding (longitude, latitude) coordinates are needed to identify 95 % of all users. For the social dataset, the figure shows that at least 55 points on average is needed to identify 95 % of all users. This amount of points is much larger than for the results given by de Montjoye et al. and for the **Home/Work** labels seen in Figure 4.9, also indicating that social labels are more privacy preserving than spatial labels when considering two labels. Furthermore, the green bar shows that there is some deviation from the mean of the 95th percentile to the minimum number of points required to uniquely identify 95 % of all users. For the **Home/Work** labels, the points differed less, as seen in Figure 4.9, but the for the social labels the points deviate some. This may be because of the shape of the distribution of the 95th percentiles as seen in Figure 5.7. If the distribution was normal, taking the mean would then give the *typical* amount of points to identify a user 95 % of the time, but as the distribution is strangely shaped as in Figure 5.7, the 95th percentiles become more uncertain. This observation can also be seen when

looking at the large spread of the standard deviations in Figure 5.9.
In the following section, the results from the temporal correlation analysis performed on the **dialogue** dataset will be presented.

5.3 Temporal correlations

As discussed in Section 5.2, it is presumably harder to identify users based on their social activities rather than their spatial activities. To verify this, the temporal correlations for the unicity simulations for the **dialogue** dataset have been found. The results for the daily and weekly resolutions can be seen in Figure 5.10, where the blue curve are the temporal correlations and the cyan (light blue) curve is the *density curve*. The density curve is the curve $1 - \rho(1 - \rho)$, where ρ is the density. As a reminder, the density is the probability of observing a 1 at the given time. A full explanation on this curve is given in Section 3.5.

In the left figure in Figure 5.10, one can see the results for a daily resolution, which is the mean fraction of users discarded at each iteration for all users. Looking at the left figure in Figure 5.10, it seems like not many users are being discarded at each iteration. To remind the reader, a value of 1 on the second axis means that no users have been discarded and a value of 0 means all users have been discarded. Comparing the results for the daily resolution seen in Figure 5.10 with the results for the **Home/Work** dataset seen in Section 4.1, the values on the second axis are generally higher for the mobility dataset, ranging between approx. 0.75 and 0.975, than for the **Home/Work** dataset, which has the largest value

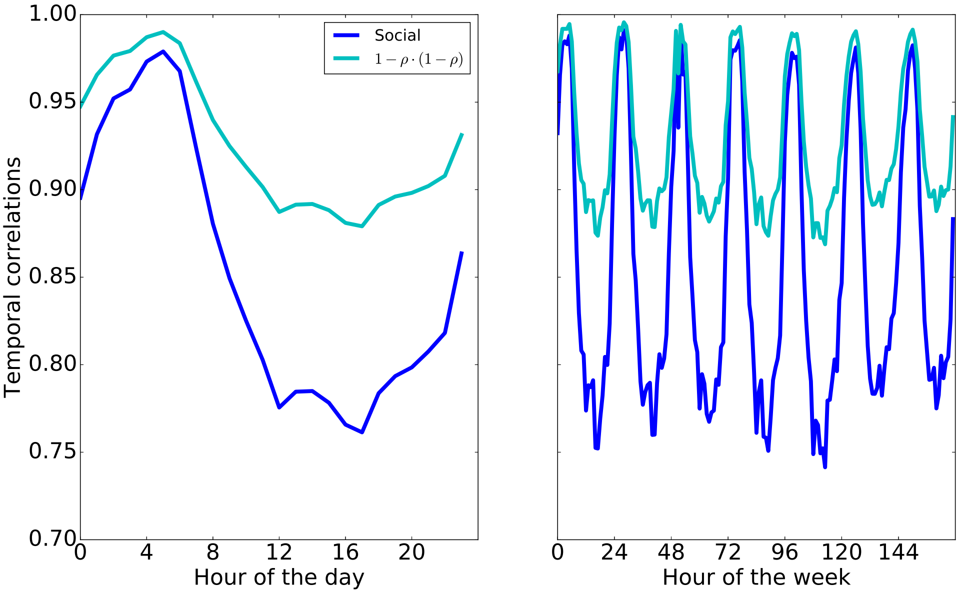

Figure 5.10: *Figure showing the mean temporal correlations for the **dialogue** dataset. The blue curve is the temporal correlation curve and the cyan (light blue) curve is the so-called density curve. **Left figure:** Figure showing the results from the temporal correlation analysis for a daily resolution. Users are easier to find when using time bins that represent the middle of the day rather using the night hours. **Right figure:** Figure showing the results from the temporal correlation analysis for a weekly resolution. The results for the weekly resolution is the same pattern as for the daily resolution, where only a small change happens in the middle of the weekends. Comparing these results with the **Home/Work** dataset, users are in general harder to find, also as discussed in Section 5.2. Both the daily and weekly curves follow the pattern of the density curve, which is expected, however with some deviation in the daily hours. The small span on the second axis is noticeable, yielding only small changes in temporal correlation.*

around 0.81 for Scenario 2. This behaviour also corresponds to the behaviour seen in Section 5.2. From the left figure in Figure 5.10 it is also evident that users are easier to find when using bins that represent the middle of the day and harder to find when using bins that are representing the evening and night hours. Additionally. the temporal correlation curve follows the pattern of the density curve as expected. There is however some deviation in the middle of the day, but the general trends are the same.

In the right figure in Figure 5.10 one can see the mean temporal correlations for all users with a weekly resolution. From the weekly resolution in Figure 5.10 the same pattern emerges as for the daily resolution, that is it is easier to identify users when using bins from the middle of the day rather than from the evening or night hours. It can be seen that this result is given for all days, so using either day of the week may not change the results drastically. However, it can be seen that it is slightly harder to find users when using bins in the middle of the day for Saturday and Sunday rather than the weekdays, however not much. Again, the values on the second axis are higher than for the mobility dataset, further stating the claim that social observations yields harder identification of users. As for the daily resolution, the pattern for the temporal correlation curve and the density curve are similar, however there are deviations when comparing them in the daily hours.

It should be noted however that there is a skewness in the number of time stamps. They are not evenly distributed over the daily course, which can be seen in Figure 5.3. In the figure it can be seen that there are much fewer observations in the early morning and night hours of the day, which is caused by social events only being recorded when they happen. Therefore, the time stamps that are being used for the identification of users are mostly in the hours where the temporal correlation shows that many users are being discarded. These time stamps are therefore more common than time stamps which discard fewer users, and the distributions of the 95th percentiles may in fact be different. A way to come around this *problem* may be to infer missing data which then gives an evenly distributed number of time stamps and then use the scenarios from the attack model, but it also worsens data quality to give users more missing data. Another way could be to infer an *alone* label in the missing data, as this is the most probable label when not knowing which other labels to consider. However, this has not been done since it would require more work which the time limit did not permit and may not change the results drastically.

To see the individual importance of the temporal correlations for users with different findability, these temporal correlations have been found for both a daily and weekly resolution. Users who have a high, medium and low findability have been chosen, where the user with a high findability has the minimum 95th percentile, the user with medium findability has the median 95th percentile and the user with low findability has the maximum 95th percentile, as given by the results in Section 5.2. The temporal correlations for a daily resolution for users with a high, medium and low findability can be seen in Figure 5.11 together with the density curve.

From Figure 5.11 one can see that the higher the findability, the lower the curves, especially in the daily hours. In the night hours, all users have very high temporal correlation values, which is because the number of time stamps is limited as opposed to the number of time stamps in the daily hours, as can be seen in Figure 5.3. This observation yields that almost no information is given in the night time bins, which is also because users are usually much less active in these hours of the day. Intuitively however, if a user has a social event in the night hours where no other user has, the user who has the event would be easy to find, but since so few observations are in the night hours it is not evident in the temporal correlation curves. The user with high findability has the lowest curve, which is because a lot of users are discarded at each iteration, where the user with low findability has the highest curve because of few discarded users at each iteration. The curve for a user with low findability follows the density curve more than the users with medium and high findability, respectively. It can also be seen in the figure that

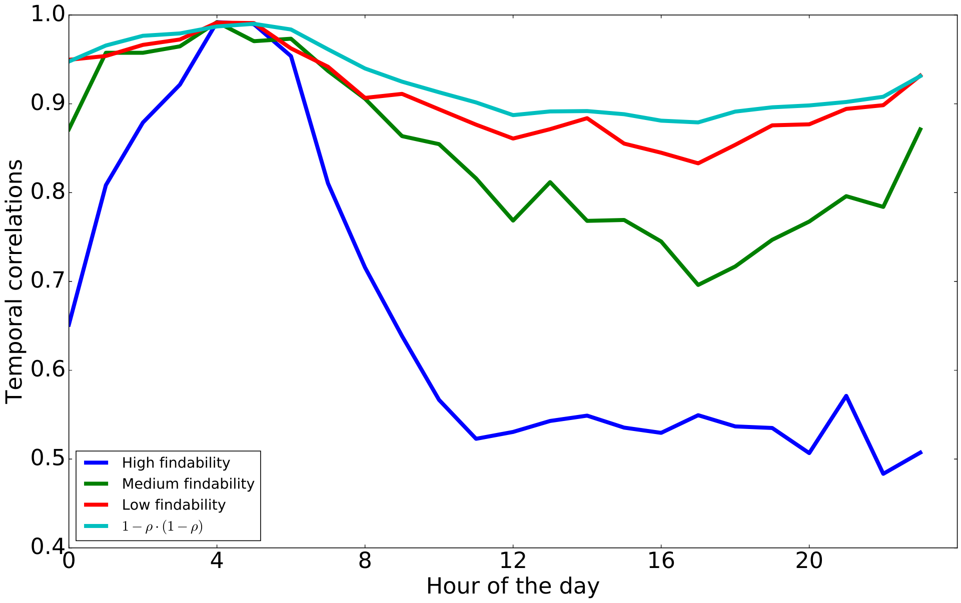

Figure 5.11: *Figure showing the temporal correlations for users with different findability on a daily basis. It can be seen that the higher the findability, the lower the curves and especially in the daily hours. The curve for a user with low findability follows the density curve more than for users with a medium and high findability, where the user with the high findability has the lowest temporal correlation values. The curves follow the density curve, however with deviations in the daily hours.*

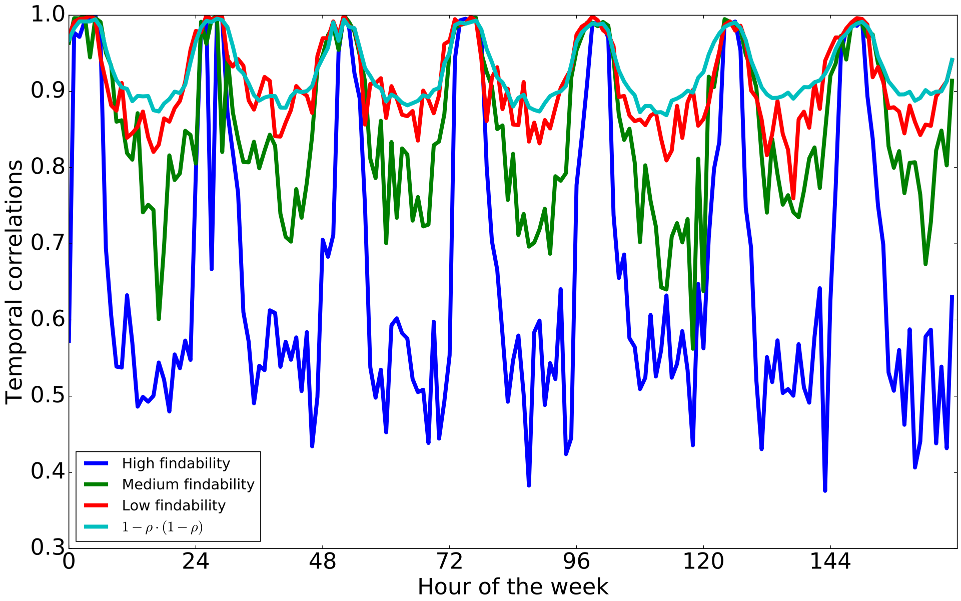

Figure 5.12: *Figure showing the temporal correlations for different users on a weekly basis. It is given that the higher the findability, the lower the temporal correlation values are in the daily hours. It can also be seen that almost no information is given in the nightly hours, since all curves have the same values. The curves also follow the density curve, but with large deviations in the daily hours.*

the user who has the highest findability can easily be found with time bins in the daily hours and not so much in the night hours.

In Figure 5.12, one can see the temporal correlations for the three different users in a weekly resolution. It can be seen that the user who has the highest findability has more temporal information where the other two users has less. This is kind of the other way around as opposed to the **Home/Work** dataset, but again it may be because there are more time bins in the night hours than the day hours, and since a user with high findability will discard more users at each iteration, the curve has lower values in the time bins that have been used. From the figure it is also evident that there is the same pattern as for the daily resolution, and that the curves follow the density curve, but with large deviations in the daily hours.

In this chapter identification of users using social labels has been discussed. The social datasets described in Section 2.3 have been combined into a single dataset called the **dialogue** dataset to improve performance. The results seen in both Section 5.2 and 5.3 indicate that using social labels have a more privacy-respecting labelling than the mobility labelling of being at home or not as seen in Chapter 4. Looking at the temporal correlations for users with different findability also yielded information that the higher the findability, the more information is given in the daily time bins. Also, almost no information is given in the nightly time bins, since the different users almost have the same values here, as seen in Figures 5.11 and 5.12. This is caused by the fact that there are much fewer time bins in these hours of the day and because users usually sleep in these hours and are not as socially active.

Chapter 6

Four spatial labels

In this chapter an analysis performed on the four label mobility dataset described in Section 2.4 will be considered. In Section 6.1 the results from the unicity simulations when applying the attack model from Section 3.1 will be discussed. Furthermore, in Section 6.2 the temporal correlation analysis described in Section 3.5 will be performed. To get insight in how much the spatial resolution impacts the findability of users, an analysis on the top three locations of users will be performed such that results can be compared to the findability of the top one location which was done in Chapter 4. Intuitively, users should be easier to identify, since more information is included in the data yielding higher-dimensional data and users are more unique with this information.

6.1 Unicity

In this section the results obtained from using the attack model described in Section 3.1 applied on the four labels spatial dataset will be discussed. Box plots similar to the box plots seen in Chapter 4 have not been produced, as the data quality is close to being the same for users in this dataset and may not affect the convergence of identifications for the scenarios depending on users' data quality.
In Figure 6.1 one can see the distribution of the 95th percentiles for all users for Scenario 1. Users with less than 10 % valid data have been removed since these may not contribute much to the unicity and also such that comparisons can be made to the results obtained from the **Home/Work** dataset. From Figure 6.1 one can see that a lot of users have been found around ten points, with a long tail after ten

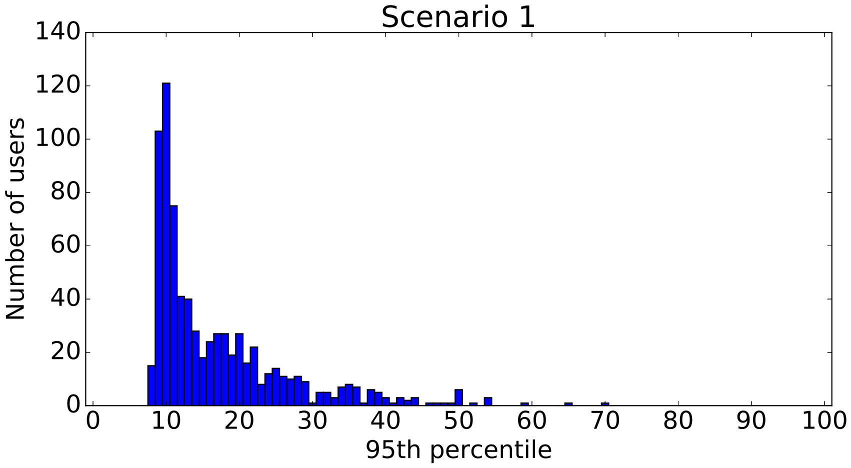

Figure 6.1: *Figure showing the distribution of 95th percentiles for all users for Scenario 1. Here users with less than 10 % non-missing data have been removed. A lot of users have been identified after ten points, with a long tail occurring after ten points. Furthermore, comparing this figure to Figure 4.3 it is easier to identify users using four labels instead of two.*

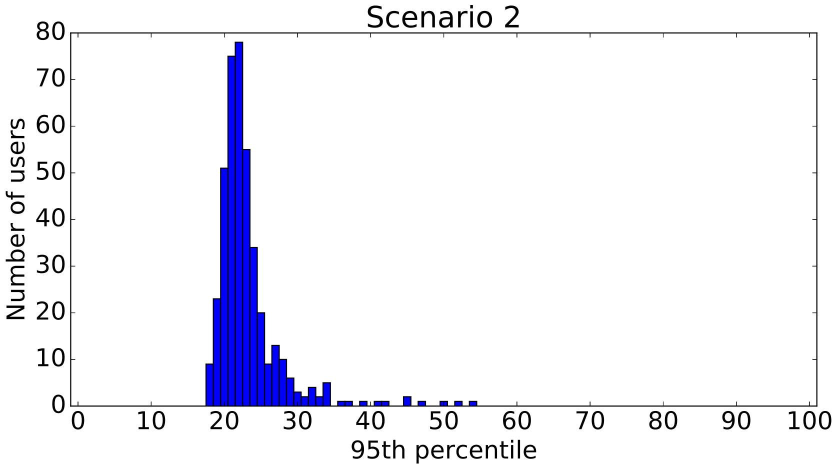

Figure 6.2: *Figure showing the distribution of 95th percentiles for all users for Scenario 2. Here users with less than 50 % non-missing data have been removed. The number of users left is 410, where it was 413 for the* **Home/Work** *dataset, but the comparison can still be made, since the difference is not that large. Comparing the figure to Figure 4.5 yields an easier identification when using four labels instead of two, as was the case for Scenario 1.*

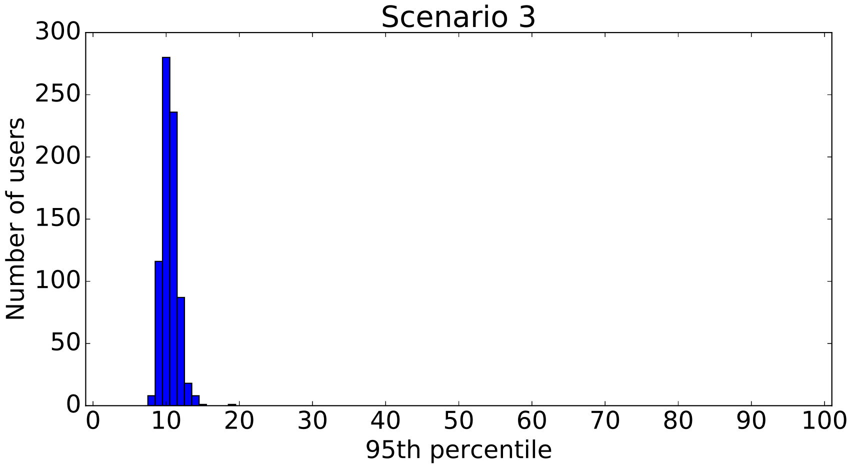

Figure 6.3: *Figure showing the distribution of 95th percentiles for all users for Scenario 3. Here users with less than 10 % non-missing data have been removed, which is the same for the two labels dataset. Comparing the figure to Figure 4.7 gives that it is much easier to identify users with four labels instead of two in this scenario, as the distribution is much more narrow and shifted towards zero than for the two labels.*

points. It seems as though that users are quite different using this scenario, which was the same case as for using the **Home/Work** labels in Chapter 4. Furthermore, comparing Figure 6.1 with Figure 4.3 for the two labels, it can be seen that it is generally easier to identify users for the four labels spatial dataset. It should be stated that the number of users after removing users with 10 % valid data are the same for the two labels and four labels datasets, which is 755 users, so the comparison is fully valid. The two distributions are very similar, but when using four labels instead of two, the distribution is shifted a few points towards zero. This is also in correspondence with the intuition that it is easier to identify users using more labels.

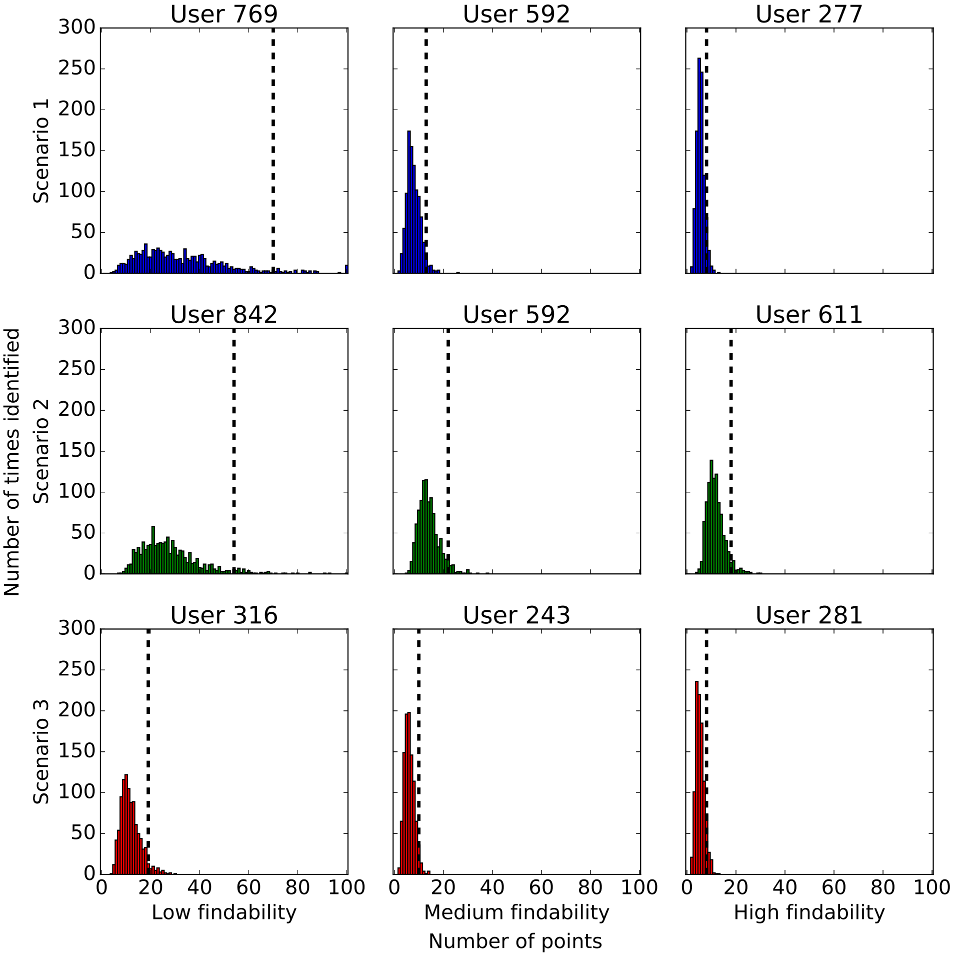

Figure 6.4: *Figure showing histograms of identifications of different users. Histograms are shown for all scenarios for users with low, medium and high findability, where the black, dashed lines show their 95th percentiles. As for the* ***Home/Work*** *labels, moving upwards in findability yields a smaller spread in the number of points needed to identify a person. Comparing the histograms with the histograms in Figure 4.8, one can see that for all users there is less spread in the distributions, especially for users with low findability, meaning easier identification when using four labels instead of two.*

In Figure 6.2 one can see the distribution of the 95th percentiles for Scenario 2. Here users with less than 50 % valid data have been removed, since not all users can be found if more are kept. The number of users left are 410, where it was 413 for the **Home/Work** dataset, so they differ slightly. This should be kept in mind when comparing them, but the difference is not large so the comparison is still valid. From Figure 6.2 one can see that all users have a 95th percentile less than 100, and that approximately half of the users are found after 22 points. After 22 points there is a short tail with some users being more hard to find. Comparing these results with the results for the two labels spatial dataset, which can be seen in Figure 4.5, one can see that it is easier to find users with four labels than with two.
Finally, in Figure 6.3 one can see the distribution of 95th percentiles for Scenario 3. As for Scenario 1, users with less than 10 % valid data have been removed, again yielding the same number of users as for the **Home/Work** dataset, which makes it easy to compare the two.

Looking at Figure 6.3 the distribution almost seems normal, which means that around half of the users are identified around the peak, which is approximately 10 points. Comparing these results with the results for the **Home/Work** dataset, the identification is much easier for four labels than two for Scenario 3, as the distribution in Figure 6.3 is much more narrow and shifted towards zero than the distribution seen in Figure 4.7. Comparing all scenarios in this section yields the hardest identification of users with Scenario 2, with the easiest identification of users for Scenario 3. This was also the case for the two labels dataset so the scenarios are not affected by the spatial resolution of the data.

When identifying users, sometimes there is much variation in the number of points needed to identify them. This was seen in Figure 4.8 for the **Home/Work** dataset, where users with low findability had much more variation than users with high findability. Such a figure for the four labels spatial dataset can be seen in Figure 6.4. In the figure, histograms are presented for users with different findability and are showing the number of identifications for different number of points. Low findability is the maximum 95th percentile, medium findability is the median 95th percentile and high findability is the minimum 95th percentile. Also, users' 95th percentiles are presented as black, dashed lines. Looking at Figure 6.4, one can see that again there is much more variation in the number of points needed to identify a user for users with low findability than for users with high findability. Comparing Figure 6.4 to Figure 4.8, one can see however that fewer points in general are needed to identify users when using four labels instead of two, which also could be seen in the histograms for the 95th percentiles. Also, much less spread is present in the histograms when considering four labels over two, especially for users with low findability, which makes sense since increasing the resolution yields users easier to *stand out* and therefore easier identifications.

To compare the results from the four labels spatial dataset with the results in the literature, a figure showing the mean fraction of unique users after a certain amount of points used can be seen in Figure 6.5. This figure shows the same results as the results found by de Montjoye et al. [14], and is similar to Figure 4.9 for the **Home/Work** labels and Figure 5.9 for the **dialogue** labels. In the figure the green bars are the average 95th percentile for all scenarios and the red bars show the spread in the standard deviation of the distributions. Figure 6.5 shows that the minimum number of points needed to uniquely identify 95 % of all users is given for Scenario 3 and is around 11 points. This is less than for both the **Home/Work** labels, which can be seen in Figure 4.9 to be around 15 and for the **dialogue** labels as seen to be 53 points in Figure 5.9. The width of the distribution of the 95th percentiles is shown to be the narrowest for Scenario 3 and the broadest for Scenario 1. This spread for Scenario 1 can be seen in Figure 6.1, and the shape of the distribution of the 95th percentiles may cause the mean 95th percentile to deviate some from the minimum number of points needed to identify at least 95 % of all users.

For Scenarios 2 and 3 the mean 95th percentiles are close to the mean minimum number of points needed to identify 95 % of all users, so the results of the 95th percentiles are in correspondence with the work done by de Montjoye et al. However, the 95th percentiles are to be preferred over the simple aggregation of users, since each user has a certain findability which can be used on an individual level. Again, de Montjoye et al. reports that 4 spatio-temporal points are enough to identify 95 % of all users, but more is needed here as higher-dimensional data is used in the literature over the labels used here, which makes sense.

In this section the results from the unicity simulations for the four labels spatial dataset was compared to the results for the two labels spatial dataset. It can generally be seen that using four labels instead of two yields an easier identification for all scenarios, which also was the intuition. Furthermore, using Scenario 2 was the hardest to identify users with, and Scenario 3 was the easiest, as for the two labels dataset. Comparing the results with the results in the literature also yields harder identification when using four labels instead of raw GPS-coordinates. In the following section the results from the temporal correlation analysis for the four labels dataset will be discussed.

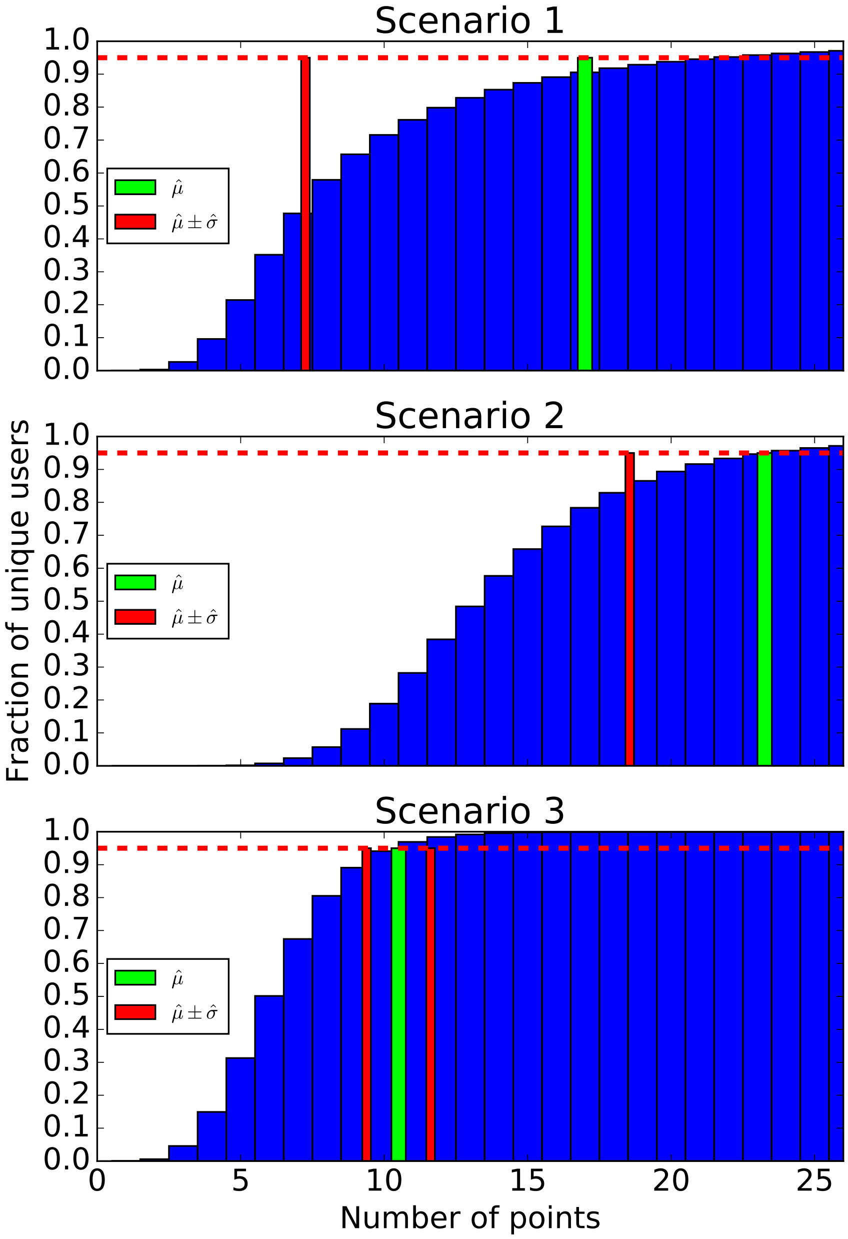

Figure 6.5: *Figure showing the fraction of users identified at each number of points for the three scenarios. The green bars show the mean 95th percentile for each scenario and the red bars show the spread of the standard deviation from the mean. The values of the mean 95th percentiles are almost the same as the mean number of points needed to identify 95 % of all users, however there is some deviation for Scenario 1. It can be seen that the minimum number of points needed to identify 95 % of all users on average is around 11 points, which is seen for Scenario 3.*

6.2 Temporal correlations

In this section the results from the temporal correlation analysis will be outlined. The temporal correlation analysis will give insight in what time bins are the most informal when users should be identified. The results in this section will be compared to the results for the two labels dataset to see the difference when using different spatial resolutions.

In Figure 6.6, one can see the temporal correlations of the unicity simulations for the four labels spatial dataset. For all users the mean has been taken in every hour of the day or hour of the week. To remind the reader, a value of 1 on the second axis means no users were discarded at the point and a value of 0 means all users were discarded at the point. Looking at the left figure in Figure 6.6, one can see that for the daily resolution not at lot of importance in the time bins are present. However, it is evident that for all scenarios when using time bins around noon, where everyone goes to lunch, it is harder to identify users than just before or after. This may be caused by users going either home for lunch and staying there or maybe first going to the university after lunch.

Looking at the right figure in Figure 6.6 one can see the weekly resolution of the temporal correlations. Here one can see that there is a general pattern that all scenarios follow more or less on the weekdays, where a special case arises in the weekends. In the weekends, especially Scenario 3 is impacted and all scenarios seem to slightly increase the chance of identifying users. Comparing the figure to Figure 4.25 for the two labels dataset it can be seen that the curves for the four labels dataset are closer towards zero on the second axis, further stating that it is easier to identify users with more labels than two. Also, when comparing the correlations for the weekly resolution, the peaks in the night hours for the two

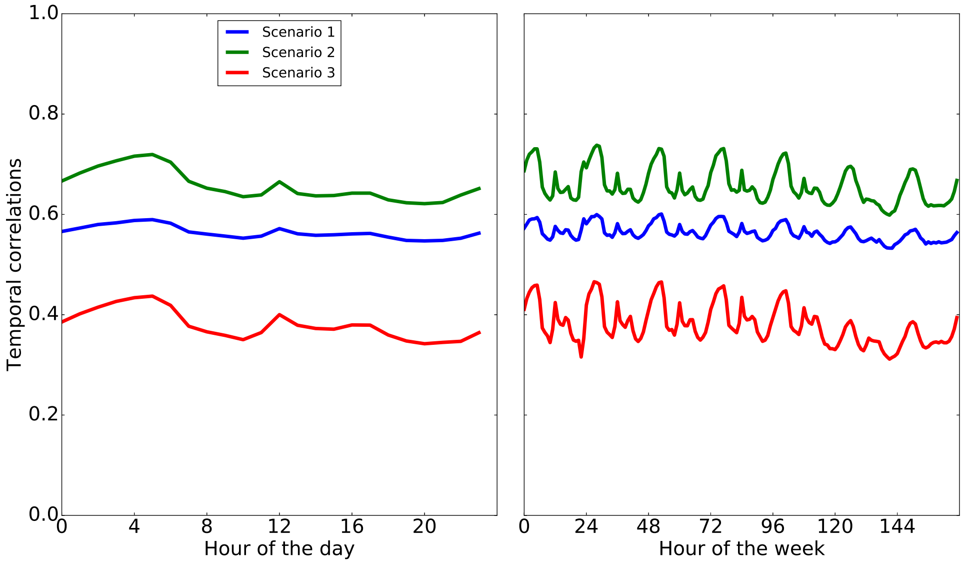

Figure 6.6: *Figure showing the temporal correlations for all users. Here the mean has been taken. The curves are lower for the four labels dataset than for the* ***Home/Work*** *dataset, which can be seen in Figure 4.25.* ***Left figure:*** *Figure showing the temporal correlations for a daily resolution. It can be seen that the curves do not change a lot, but there is some importance at specific times of the day. For example, users are easier to find when using bins before or after lunch, but using bins in the lunch time or during the night it is harder.* ***Right figure:*** *Figure showing the temporal correlations for a weekly resolution. There is a general pattern for all scenarios in the weekdays with a change in the weekends, especially for Scenario 3.*

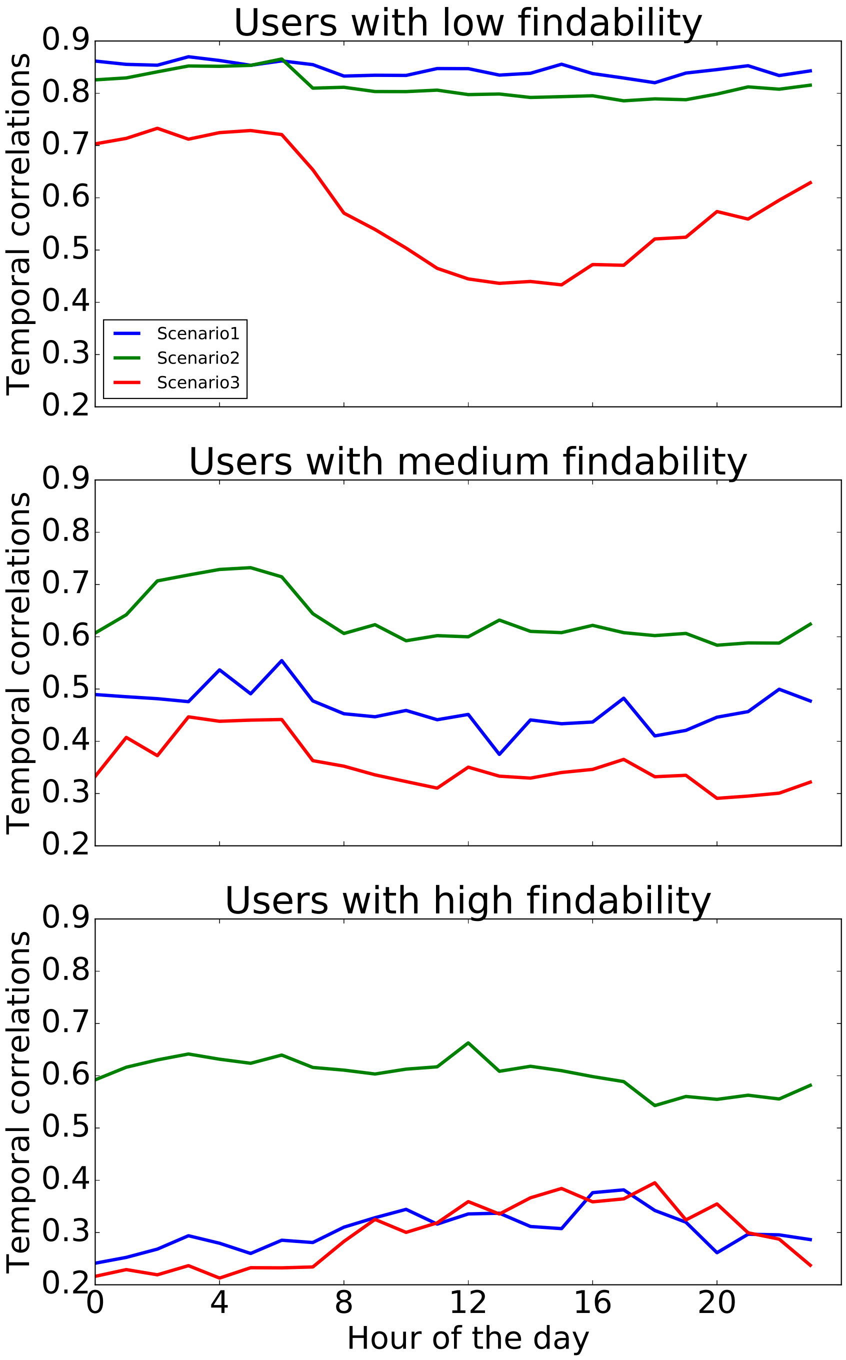

Figure 6.7: *Figure showing the temporal correlations for different users on a daily basis. It can be seen that the most information is given for users with low findability for Scenario 3 and not much information is given otherwise. The higher the findability the less the values are on the second axis, which makes sense since more users are discarded when considering users who are easy to find.*

labels dataset is higher than for the four labels dataset, yielding it easier to detect users in the night time when using more than two labels.

As for the **Home/Work** and **dialogue** datasets, an individual inspection of the temporal correlations will be performed. Users with low, medium and high findability will be inspected, where the users with low findability has the maximum 95th percentile, medium findability is the median 95th percentile and high findability is the minimum 95th percentile. The users who have been found for the different scenarios are not the same, since the distributions of the 95th percentiles are different, as seen in Section 6.1. The analysis will show how the different findabilities impact the temporal correlations.

In Figure 6.7 one can see the temporal correlations for users with low, medium and high findability for a daily resolution. In Figure 6.7 one can see that in general the higher the findability the lower the temporal correlation values. The curves have lower values when moving up in findability and the curves become more constant meaning that less temporal information is given when users become easier to find. This was also the case for the **Home/Work** labels, as seen in Figure 4.26. As not much temporal information is given for the average user, as seen in Figure 6.6, not much is given either in Figure 6.7 even for users with low findability, only for Scenario 3. However, when looking at Figure 6.8, one can see more information is given on a weekly basis.

For users with low findability, there seems to be a pattern for scenarios 2 and 3 in Figure 6.8, where the night hours have the largest values, meaning few users are discarded, and the smallest values in the middle of the day. It can also be seen that for Scenario 1 there is not a specific pattern for a user with low findability. Moving up in findability it seems, for all scenarios, that less information is given since the curves do not follow a specific pattern and are more *chaotic*. The curves are also lowered, meaning that more users in general are discarded when considering users with higher findability, which makes sense. Comparing these results with the results for the **Home/Work** dataset yields that in general less temporal information is given when considering more labels. This may be caused by the fact that users are more unique and in general are easier to find, which makes the time bins more obsolete. If you really stand out, for example by being somewhere else than home in the middle of the night, you are easy to find and having more labels than two makes this easy.

In this chapter an analysis was performed on a four labels spatial dataset where users were identified using the attack model described in Section 3.1. The distributions of the 95th percentiles as seen in Section 6.1 yielded that four labels makes it easier to find users rather than two when considering spatial labels. When analysing the temporal correlations, not as much information in the time bins are given as for the **Home/Work** labels as seen in Figures 6.7 and 6.8, possibly because users are more unique when considering more labels. Therefore, it is easier to stand out in any given time of day, and the information in the time bins is less important when identifying users.

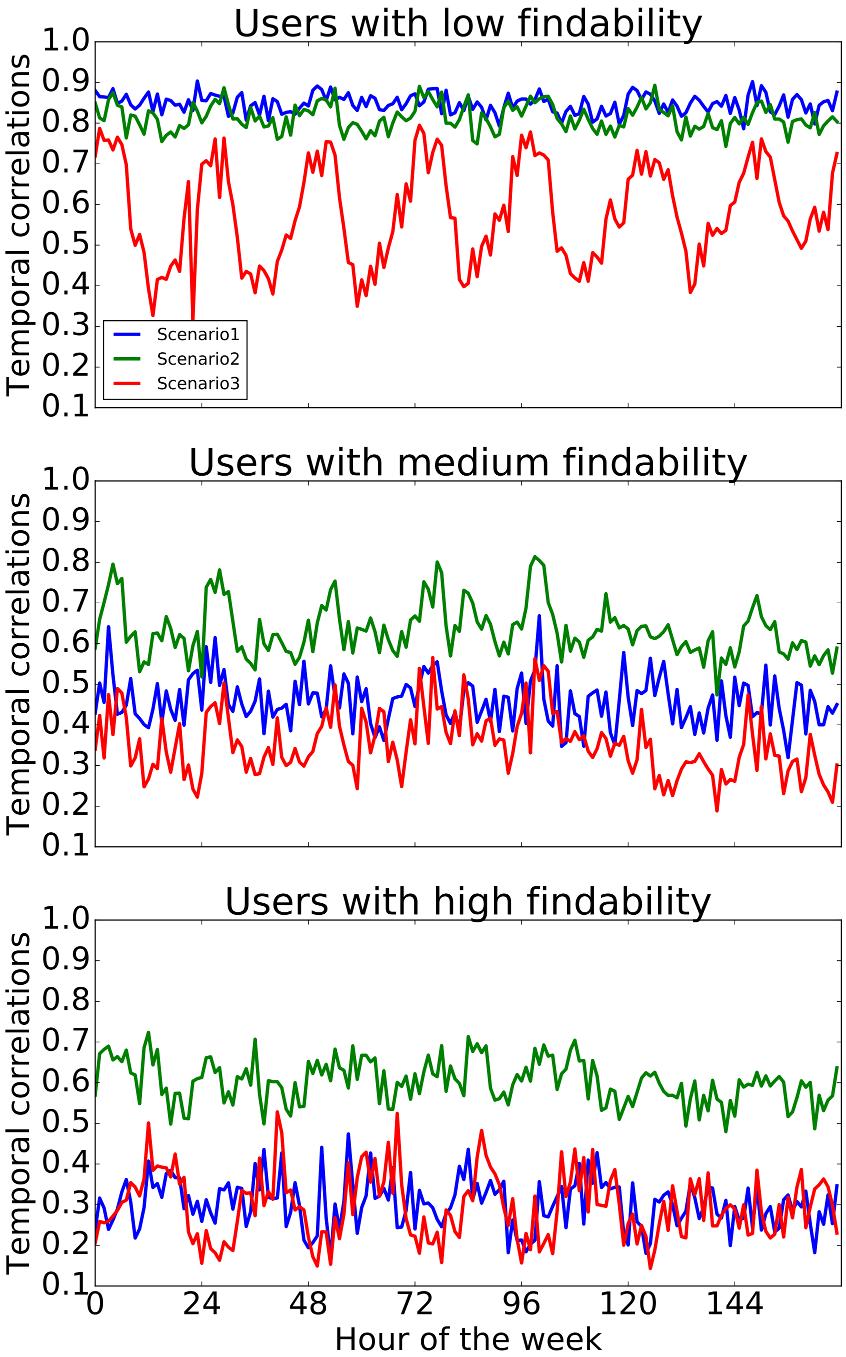

Figure 6.8: *Figure showing the temporal correlations for different users on a weekly basis. For users with low findability, there is a pattern for scenarios 2 and 3, but not as much for Scenario 1. When moving up in findability, the curves lower their values, since more users are discarded at each iteration for users who are easy to find. As for the daily resolution, generally not much information is given for in time bins.*

Chapter 7

Combined social and spatial labels

In this chapter the combined dataset holding four labels as a combination of spatial and social labels will be discussed. First, the dataset used is elaborated in Section 7.1. Then the results from unicity simulations can be seen in Section 7.2 and the results from the temporal correlation analysis can be seen in Section 7.3. As seen in Chapter 6, four spatial labels are more privacy-preserving than two labels in a large dataset, but so far this only applies to mobility records. As social data is available in the Sensible DTU study, social labels are used to mask the data with four new labels defined in Section 7.1 to see if they are more privacy-preserving than four spatial labels.

7.1 Dataset

In this dataset four labels are considered, which are a result from a combination of social and spatial labels. The social labels in this section are defined as a 1 if the user was social and a 0 if the user is not social, where *being social* is either having a Facebook/sms/call event at a given time stamp. These social labels comes from the **dialogue** dataset described in Section 5.1. Furthermore, a spatial label of 1 means that a user was at home and a label of 0 means they were somewhere else at a given time stamp, which are the labels obtained from the **Home/Work** dataset. These new labels are defined in Table 7.1. In words, the new labels mean that if a user has label 0, the user is not at home and not social at a given time. A label of 1 means that the user is not at home and is social at the given time. A label of 2 means that the user is at home and not social and finally a label of 3 means that the user is at home and is social at the given time. Additionally, if a user has missing data in the spatial dataset, that is a -1, the combined label is then also defined as missing.

In this combined four label dataset there are 782 users and 5313 time stamps. The difference in users between this dataset and the **Home/Work** dataset is caused by different users in the **dialogue** dataset, so the common users have been found and extracted. The difference in time stamps is caused by the social labels only being recorded when an event is happening, so the time stamps may differ. In Figure 7.1 one can see the number of time stamps for every hour of the day for the combined dataset. From Figure 7.1 one can see that the number of timestamps are very different on a daily basis. This is caused by users mainly being social in the daily hours and not as much in the night hours. Therefore, the results may be skewed, which will be discussed in Section 7.3. Even though some users are missing, the number of users available should still be enough to show what is intended.

Table 7.1: *New labels defined from spatial and social labels*

New label	0	1	2	3	-1
Spatial	0	0	1	1	-1
Social	0	1	0	1	0 or 1

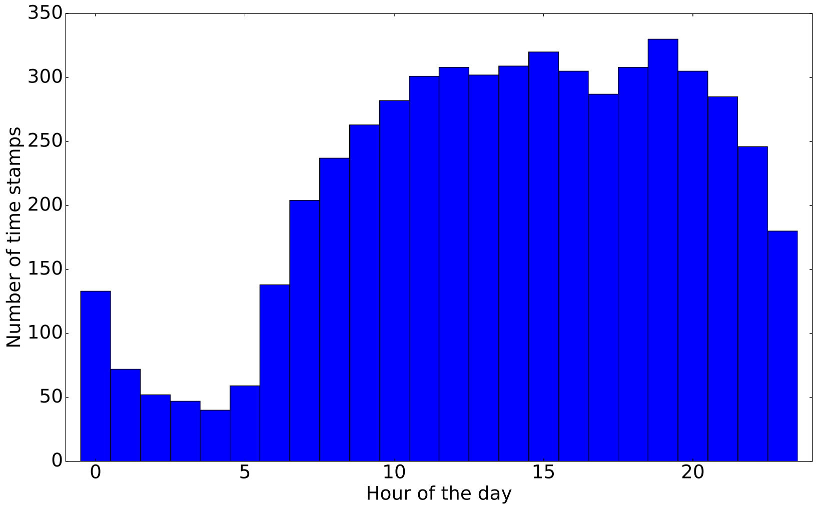

Figure 7.1: *Figure showing the distribution of the number of time stamps in the combined dataset for a daily resolution. It can be seen that the number of time stamps in the late morning and afternoon dominate the number of time stamps in the early morning and night hours.*

In Figure 7.2, one can see the fraction of usage on a daily basis for the four new labels. Here only the valid data bins have been considered. From Figure 7.2 one can see that the most used labels are labels 0 and 2 and the least used labels are labels 1 and 3, with label 3 having a order of magnitude 10 less than the others on the second axis. This may be caused by the lack of observations in the **dialogue** dataset (see Section 5.1), and therefore labels 1 and 3 are much less common. Furthermore, in general the labels have much less frequency in the mornings, and the most in the evenings. However, label 2 has a large value just before noon and again in the evenings, with a decrease between the two. It therefore seems that users that are home in the middle of the day do not socialize in these hours.

A similar figure for a weekly resolution can be seen in Figure 7.3 for all users. As seen in Figure 7.3 there seems to be a daily pattern for all labels with some change in the weekends. The most change though is for label 2, which seems more unpredictable than the other labels. Again, it can be seen that the most used labels are labels 0 and 2, and the least common labels are labels 1 and 3.
The unicity analysis performed on these labels are described in Section 7.2 and the temporal correlations are found in Section 7.3.

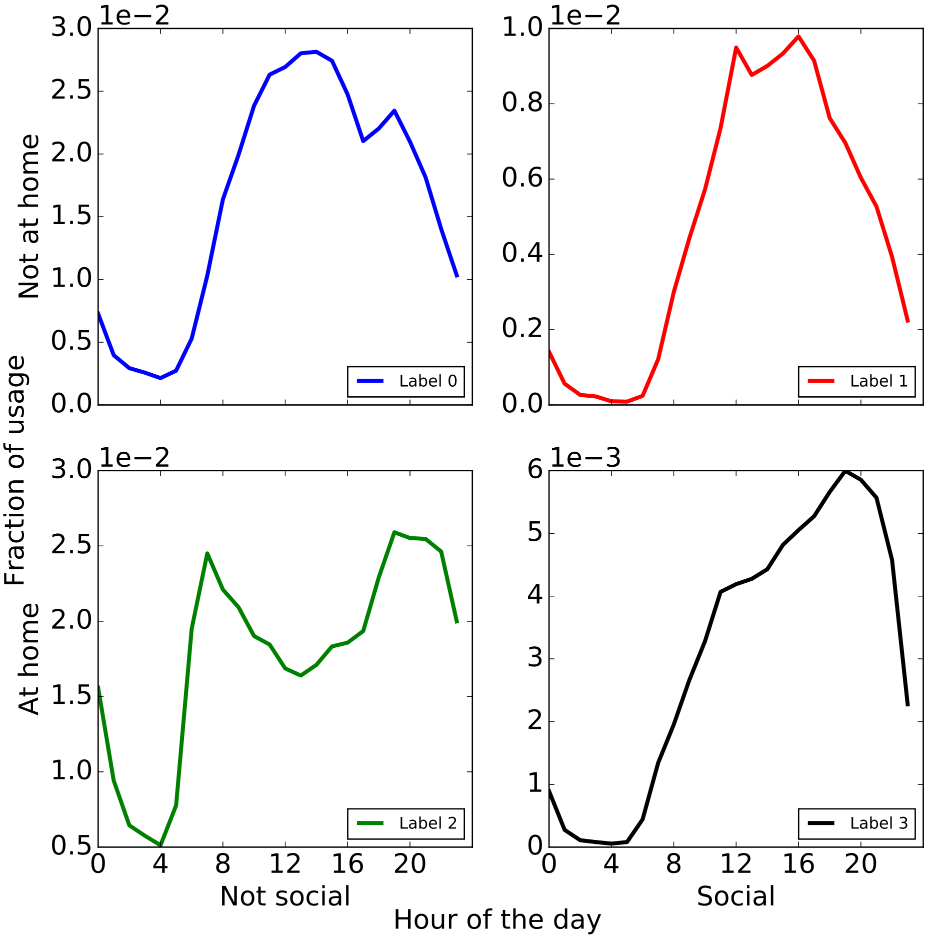

Figure 7.2: *Figure showing the fraction of usage for the labels in the combined dataset. Here a daily resolution has been chosen. The most common labels are labels 0 and 2 and the least common labels are labels 1 and 3. Label 2 seems more different than the others, which may be because not a lot of users are at home in the middle of the day and those who are may not be socially active.*

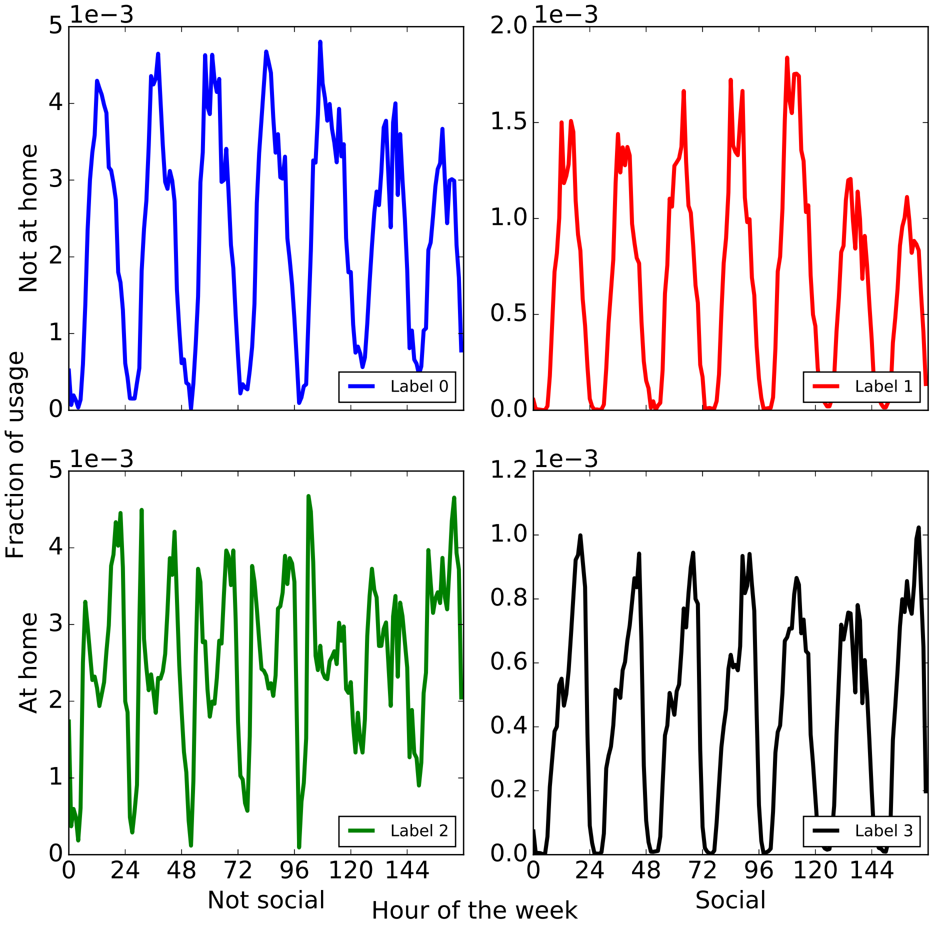

Figure 7.3: *Figure showing the fraction of usage for the labels in the combined dataset. Here a weekly resolution has been chosen. It seems as though there is a regular weekly pattern, with some change in the weekends apart from the weekdays. The most unpredictable label though is label 2 which has a somewhat chaotic behaviour, especially in the weekends. Again, the most used labels are labels 0 and 2 and the least used labels are labels 1 and 3.*

7.2 Unicity

In this section the results from the unicity tests on the combined dataset will be shown. In Figure 7.4, one can see the distribution of the 95th percentiles for all users except those with less than 10 % non-missing data. Removing these users yield 742 left, which is 13 fewer than for the four labels spatial dataset. This is caused by the different number of time stamps in the two datasets which is much fewer in the combined dataset due to the common time stamps for the **Home/Work** dataset and the **dialogue** dataset. The comparison between the combined dataset and the four labels spatial dataset is assumed to still be valid, since the difference is not that large. From Figure 7.4 it can be seen that the peak of the distribution is around 10 points, which approximately is some fewer than the half of the users since there is a tail after 10 points stretching to 48 points. Comparing the figure to Figure 6.1, which shows the same distribution for the four labels spatial dataset, gives that more users are found at the peak for the combined dataset, as well as the tail not being as long as for the spatial dataset. This gives the impression that the combined labels makes it easier to find users than for the spatial labels in Scenario 1.

In Figure 7.5 the distribution of the 95th percentiles for Scenario 2 is given for all users except those having less than 50 % non-missing data. Here there are 547 users left, which is 137 more users than for the four labels spatial dataset. This is a quite large difference in users, but the comparison should still be made although with some precautions. The number of users for the combined dataset is larger than for the spatial dataset when removing users with less than 50 % valid data since there are much fewer time stamps, where a lot of them could have had missing data. Therefore, some users may have better data quality in this dataset than for the spatial dataset.

Looking at Figure 7.5, it can be seen that there is a peak at 20 points with a small tail afterwards so approximately half of the users are found after the peak. Comparing Scenario 2 to Scenario 1, in general more points are needed to identify the users, which also was the case for the spatial datasets (both **Home/Work** and the four label dataset). However, comparing the results in Figure 7.5 to the results for the spatial dataset seen in Figure 6.2, one can see that the peak is closer towards zero for the combined dataset. This suggests, as for Scenario 1, that the combined labels do not make more privacy-preserving labels, but on the contrary make users easier to find.

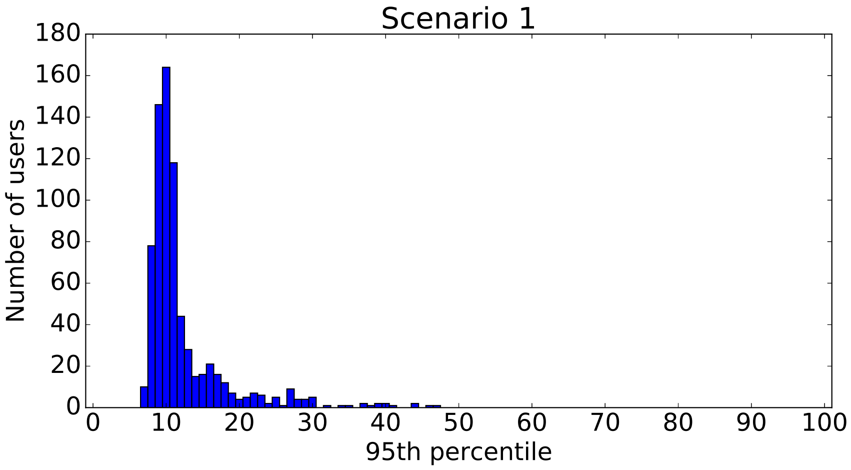

Figure 7.4: *Figure showing the distribution of 95th percentiles for all users for Scenario 1. Here users with less than 10 % non-missing data have been removed. The peak can be seen at 10 points with a tail stretching to 48 points. This tail is not as long as for the spatial four labels dataset and the peak is not as high, indicating that it is easier to identify users using combined labels instead of spatial labels.*

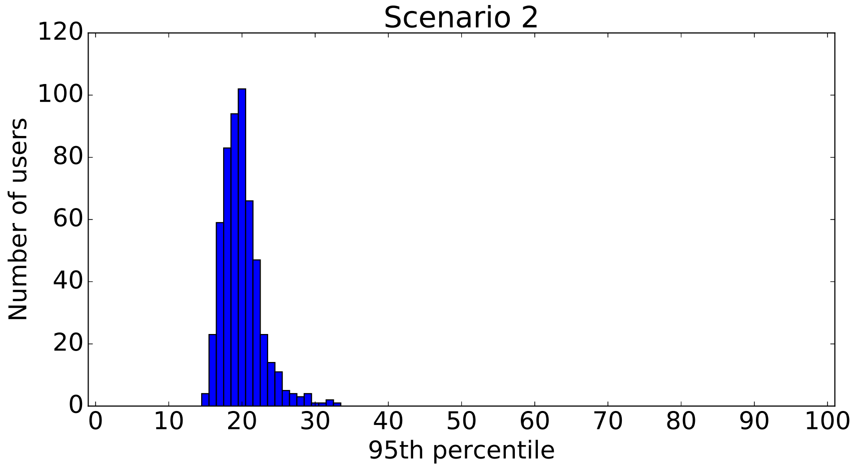

Figure 7.5: *Figure showing the distribution of 95th percentiles for all users for Scenario 1. Here users with less than 50 % non-missing data have been removed. The peak of the distribution is seen at 20 points with a small tail following. Comparing this figure with Figure 6.2, one can see that users are generally easier to find for the combined labels than for the spatial labels*

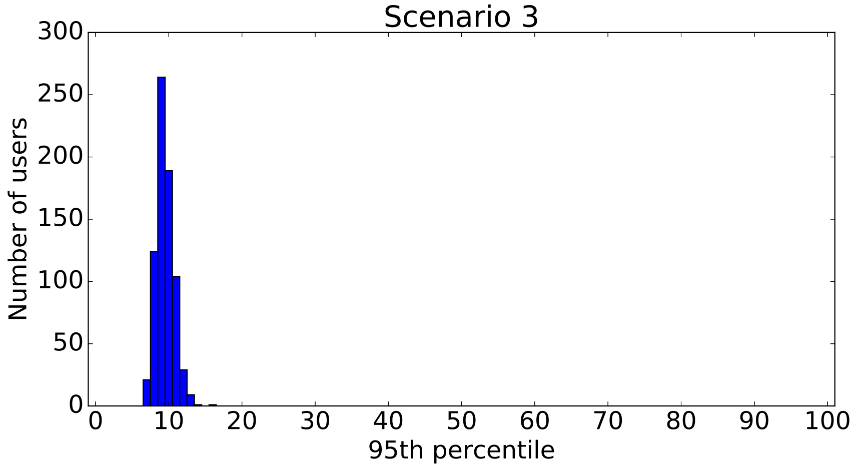

Figure 7.6: *Figure showing the distribution of 95th percentiles for all users for Scenario 3. Here users with less than 10 % non-missing data have been removed. The figure shows that around half of the users are found around 9 points where the peak is, since the distribution resembles a normal distribution.*

For Scenario 3, the distribution of the 95th percentiles for all users except those with less than 10 % non-missing data is given in Figure 7.6. From Figure 7.6 one can see a quite narrow distribution resembling a normal distribution with a peak at 9 points. Therefore, around half of the users are found after using 9 points, which is quite low. Comparing Figure 7.6 with Figure 6.3, one can see that the distributions look a lot alike, but the distribution for the combined labels has been shifted one point towards zero. Again, as for the other scenarios, this gives the information that users are easier to find for combined labels than for spatial labels when considering four labels.

Figure 7.7 shows histograms of identifications for users with different findability, with their respective 95th percentiles as black, dashed lines. To find the users with low findability, the maximum 95th percentile has been chosen, where the median and minimum has been chosen for users with medium and high findability, respectively. Comparing Figure 7.7 to Figure 6.4, it can be seen that there is less spread

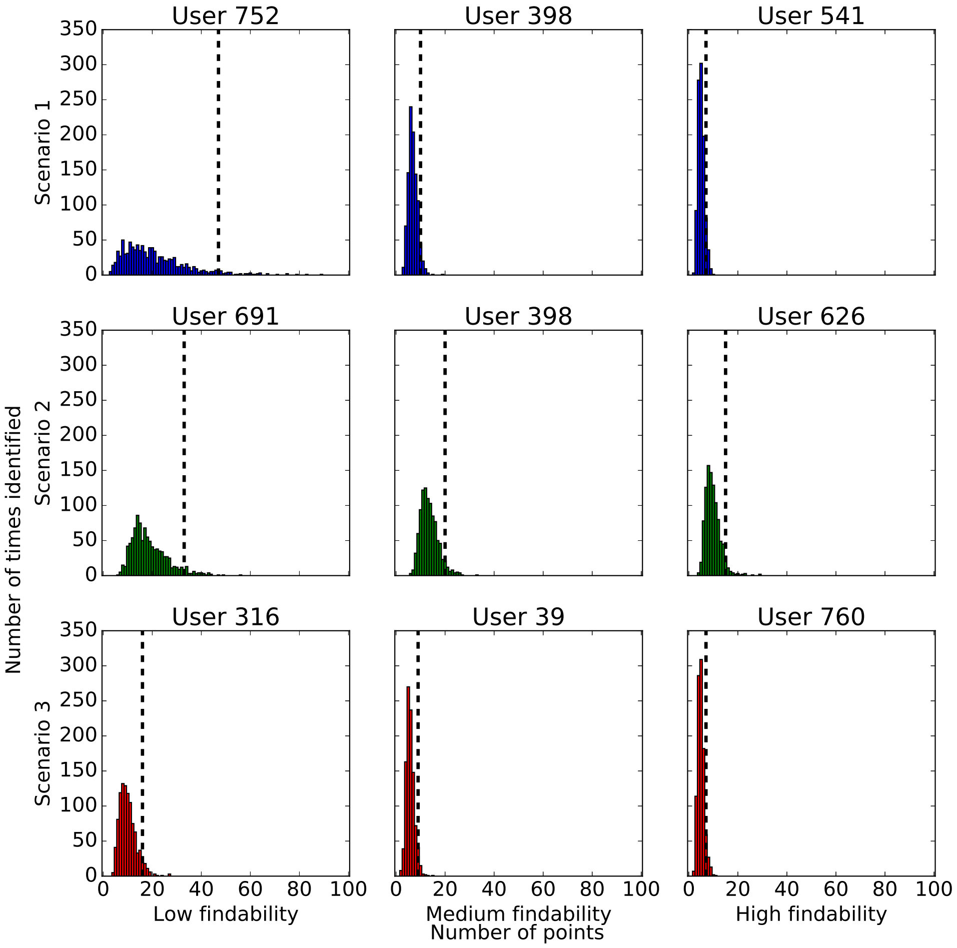

Figure 7.7: *Figure showing histograms of identifications for different types of users for all scenarios. The black, dashed lines indicate the users' 95th percentiles. Comparing this figure with Figure 6.4, one can see that there is less spread in the distributions for all users, but most noticeable for users with low findability when considering the combined labels over the four spatial labels. Although the users are not the same, a direct comparison can not be made, but the comparison it still valid, as the maximum 95th percentile has been used to find users with low findability.*

in the number of points needed to identify a user for the combined labels than for the four spatial labels. This observation is most noticeable for users with low findability, but also noticeable for the other users when looking at the number of identifications on the second axes. One should keep in mind that the users are not the same, but the comparison is still valid since the maximum 95th percentile has been chosen to identify users with low findability for both datasets.

In Figure 7.8 one can see the mean fraction of users identified as a function of the number of points over all iterations. This figure is similar to the figure provided by de Montjoye et al. and is used to compare the results for this dataset with a dataset holding high resolution spatial coordinates [14]. To brush up, de Montjoye et al. highlights that 4 spatio-temporal points are needed to identify 95 % of all users when dealing with a large number of spatio-temporal points. In Figure 7.8 it is given that the minimum number of points needed to identify 95 % is 10 points, which can be seen for Scenario 3. For the other scenarios, this number is higher for Scenario 1 and the highest for Scenario 2. The spread of the standard deviation is the largest for Scenario 1 and the most narrow for Scenario 3. This was also the case for the other datasets considered. The minimum number of points required to identify at least 95 % of users is in

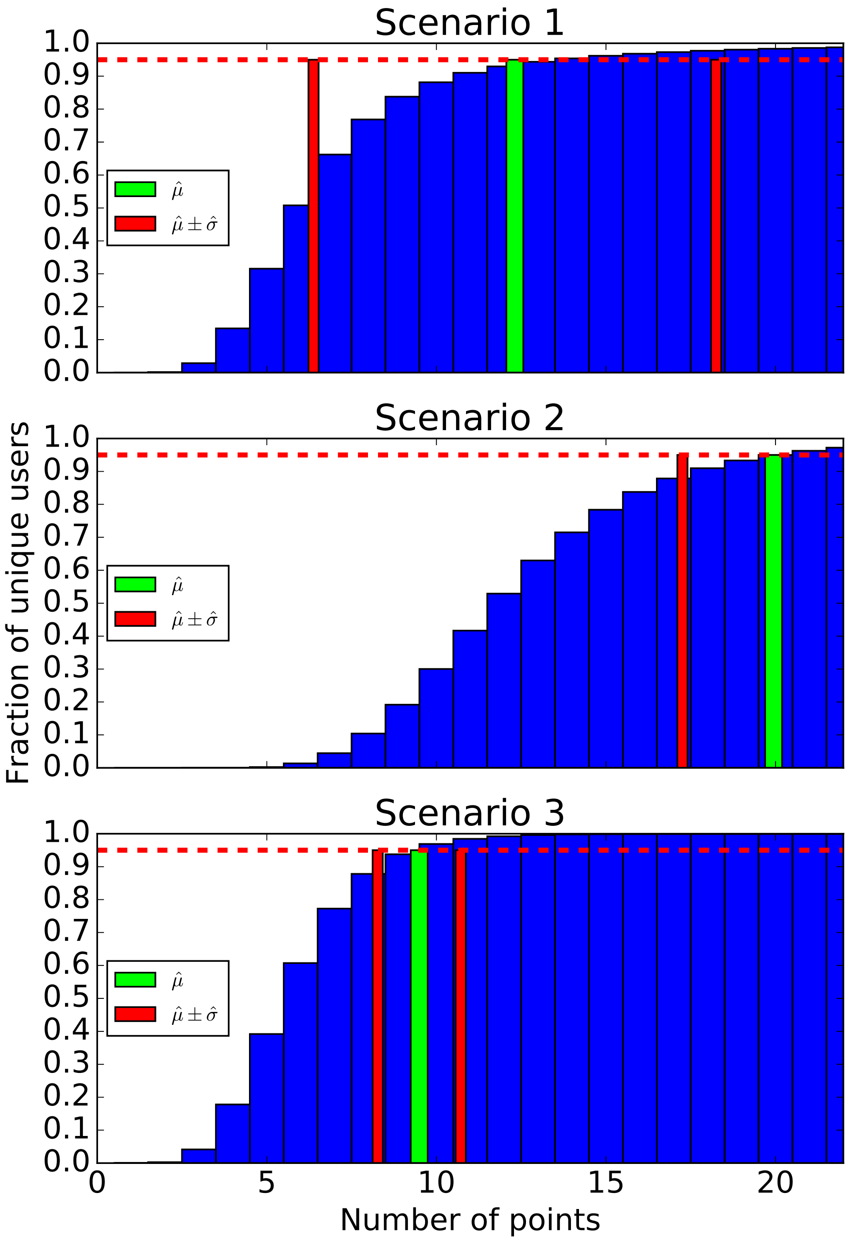

Figure 7.8: *Figure showing the mean fraction of users identified at each number of points for the three scenarios. The green bars show the mean 95th percentile for each scenario and the red bars show the spread of the standard deviation from the mean. The mean of the 95th percentiles correspond closely to the minimum number of points required to identify at least 95 % of all users.*

turn the lowest for all of the datasets considered, where it was 15 points for the **Home/Work** dataset, 53 points for the **dialogue** dataset and 11 points for the four labels spatial dataset. The mean of the 95th percentiles also correspond closely to the minimum number of points required to identify 95 % of all users as seen in the figure. Not much difference is given for the two four labels datasets however, when considering the fraction of unique users, but when looking at the distributions of the 95th percentiles for both datasets there is a difference. Indeed, a large difference can also be seen when looking at the temporal correlations, where the temporal correlations for this dataset will be elaborated in Section 7.3.

From the results in this section, it seems as the change from spatial labels to the combined labels does not change the distributions of the 95th percentile a lot when identifying users. However, the distributions for the combined labels are shifted closer towards zero, meaning that the combined labels are less privacy-preserving than the spatial labels. It therefore seems that such a masking of combining social and spatial labels does not work if one wants to hide users in a large dataset. This observation will be elaborated in Section 8.1. Also, these results may be skewed because of the available timestamps, which will be discussed in Section 7.3.

7.3 Temporal correlations

In this section the temporal correlation analysis for the combined four labels dataset will be outlined. The analysis will show what time bins hold more information than others when identifying users for the combined labels.

In Figure 7.9 the temporal correlations are shown for a daily and a weekly resolution. The temporal correlations have been found for each hour of the week and the mean has been taken for all users. Looking at the left figure in Figure 7.9, one can see the temporal correlations for the daily resolution. It can be seen that users are easier to be identified in the afternoon hours rather than in the night and morning hours. Furthermore, Scenario 2 is the hardest scenario to identify users where Scenario 3 is the easiest and Scenario 1 is in the middle. The peaks on the curve for Scenario 1 are not as noticeable as the other curves, making certain time bins more obsolete than for the other scenarios. In the right figure

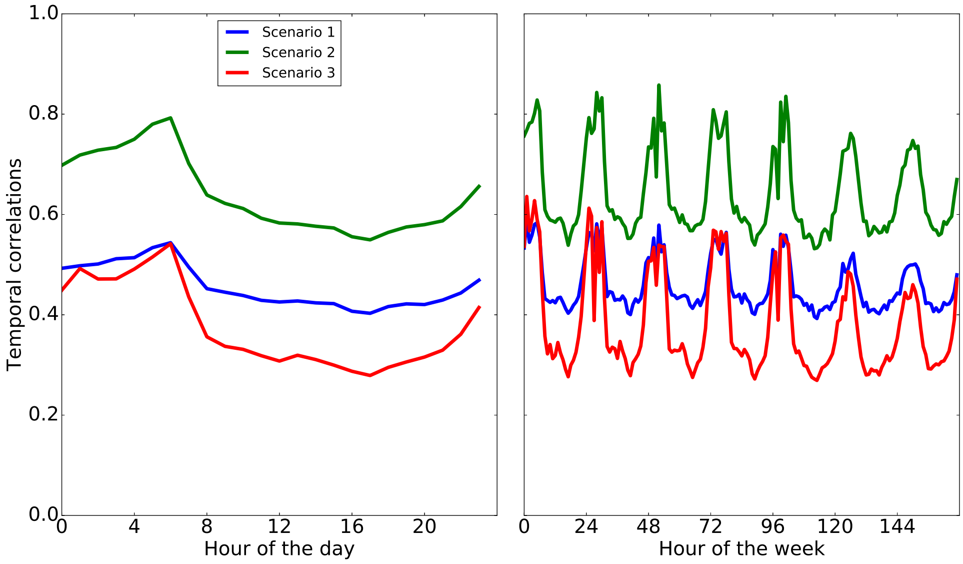

Figure 7.9: *Figure showing the temporal correlations on a daily and weekly basis for all scenarios. Here the mean has been taken for all users. Users are more identifiable during the afternoon hours and the least in the night and morning hours.* ***Left figure:*** *Figure showing the temporal correlations for all users on a daily basis.* ***Right figure:*** *Figure showing the temporal correlations for all users for a weekly resolution. The curves follow the same pattern as the daily correlations, including for the weekends.*

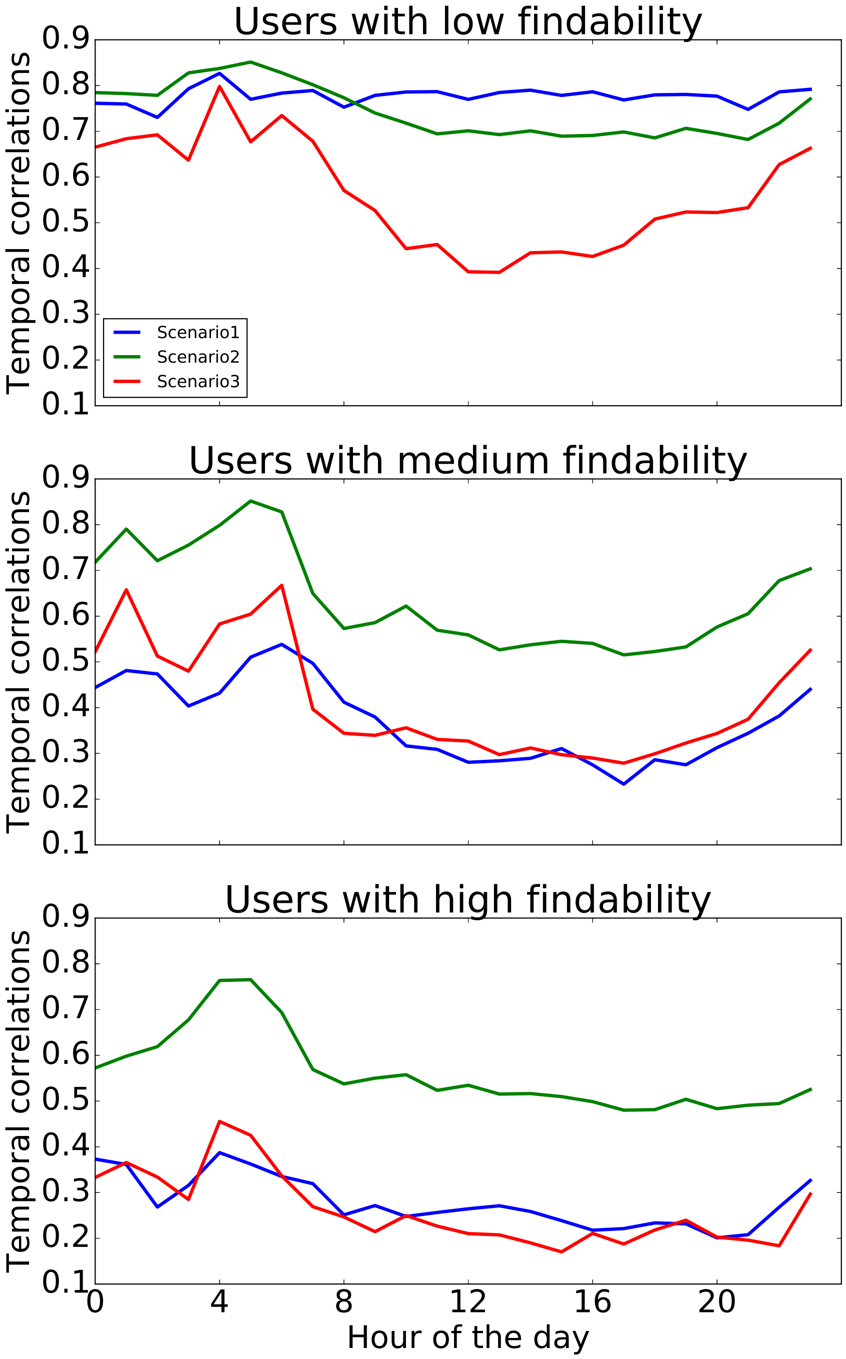

Figure 7.10: *Figure showing the temporal correlations for different users on a daily basis. Users with higher findability need less temporal information to be identified than users with low findability. The curves all have peaks in the night time no matter the findability, which is caused by users not being as socially active in the night hours than the day hours. The curves for users with high findability however are mostly flat in the daily hours.*

in Figure 7.9 the weekly temporal correlations are given. A similar pattern as the daily correlations can be seen and almost all days, including the weekends, seem to follow the same pattern. It is also evident that there is a large change from the night and morning hours to the afternoon hours.

Comparing Figure 7.9 with the results for the four labels spatial dataset seen in Figure 6.6, one can see that the curves for the daily resolution are having close to the same values, but the curves seen in Figure 7.9 are more influential on the time bins. Furthermore, for the weekly resolution there is quite a large difference from the spatial labels to the combined labels since the combined labels are deviating a lot more. These observations do not indicate that combined labels generally makes it easier to identify users rather than using spatial labels, which was the case for the 95th percentiles, but it does show that certain time bins matter much more than others. Therefore, a lot of users can be discarded when using the right time bins, even though the time bins which have high correlation values have been used earlier. As seen in Figure 7.1 one can see that there is a large number of time stamps in the late morning and afternoon hours in contrast to the early morning and night hours. Therefore, there are more time stamps in exactly the hours that yield many users to be discarded. The results are therefore somewhat skewed, and even though the results indicate that users are easier to be identified with the combined labels rather than the spatial labels, one should keep this in mind.

To inspect how the findability influences the temporal correlations, figures comparing the correlations for different types of users have been produced. In Figure 7.10 one can see the temporal correlations for users with different findabilities for a daily resolution. To remind the reader, a user who has low findability is having a high 95th percentile, a medium findability is a medium 95th percentile and a high findability is a low 95th percentile. Here the maximum, median and medium 95th percentiles have been used to find users with low, medium and high findabilities, respectively. When looking at the figure, one can see that for all kinds of users the fewest number of users is discarded in the night hours, which also is where there is the fewest number of time bins. A lot of users do not socialize in the night hours because they are sleeping, so this makes sense. Also, most users are discarded for time bins around two o'clock. It can furthermore be seen that the higher the findability, the less temporal information is given. This makes sense since if a user is hard to find, one might need *extra help* to identify him/her in the form of a time bin where the user is different from many others. There are still peaks however in the night times, but in the other times of the day the curves are mostly flat for users with high findability.

In Figure 7.11 one can see the temporal correlations for different users in a weekly resolution. From Figure 7.10 one can see that the most temporal information is given for users with low findability and the least for users with high findability, which is the general case for all datasets considered. It can be seen that the most specific pattern is for users with low findability, with high peaks in the night times and lower values in the day times. When moving up in findability, there are still peaks in the night times, but the curves become less informative in the daily hours. Looking at Figure 7.11 and comparing it to Figure 6.8 which is the temporal correlations for the four labels spatial dataset, one can see that the peaks in the night hours are not as clear for users with high findability. Again, this may be caused by the few number of timestamps in the night hours in the combined dataset, as well as users not being as socially active. For the mobility dataset, the time stamps are uniformly distributed by the hour, and users may have different habits in their whereabouts in the night hours, making them easier to find in these hours, but this is not the case for the combined dataset.

To summarize the findings in this chapter, using combined spatial and social labels makes it easier to identify people based on the analyses in this chapter. Users' 95th percentiles are lower as seen in Section 7.2 and the average number of points needed to identify 95 % of all users is lower for this dataset than for any of the other datasets. The temporal correlations however yield that when having information

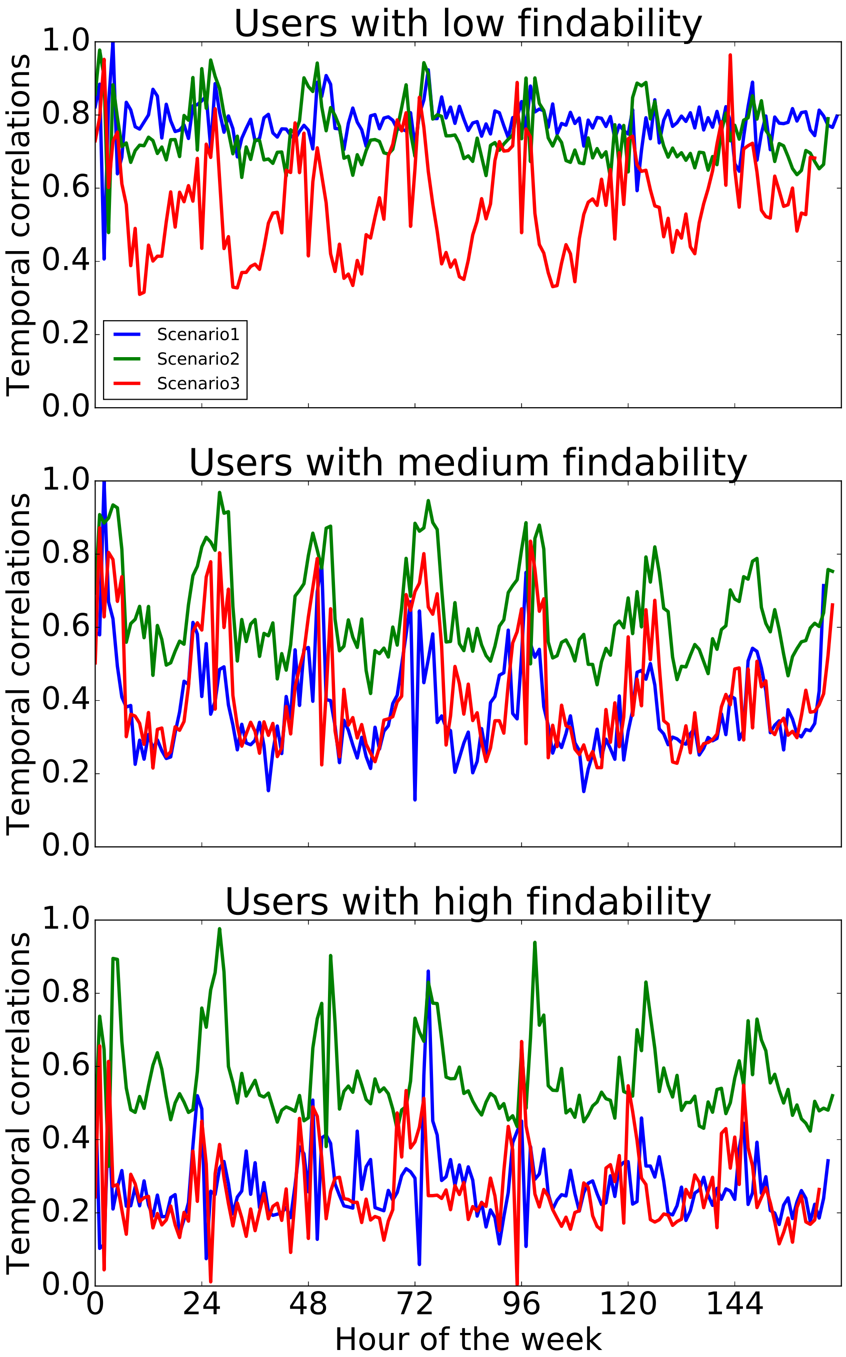

Figure 7.11: *Figure showing the temporal correlations for different users on a weekly basis. No matter the findability of users, there are high peaks in the night time which is caused by the few number of observations in these hours of the day. As the findability increases, the correlations in the daily hours decrease, which makes sense since fewer users are discarded when considering users who are easier to find than users who are hard to find.*

about any user in the night time, not many other users can be discarded, so one must have information in the daily hours to uniquely identify a person. Since there are so many time stamps in the daily hours than in the night hours for the combined dataset, this may influence the uniqueness of users, since many of the time stamps corresponding to certain hours of the day or the week may be reused when making many iterations. However, as users are not as socially active during the night, not many observations are present in such a dataset and few new observations will be in this time of the day, so the results may be fine after all.

Chapter 8 Discussion

In this chapter the results presented in the thesis as well as future work will be discussed. A discussion on the results can be seen in Section 8.1 and a section on future work can be seen in Section 8.2.

8.1 Results

The results obtained from the unicity results were acceptable and the use of different datasets made it possible to identify what made labels more privacy-preserving. As indicated in the results, using low resolutions of mobility labels was more privacy-preserving than high resolution labels, however using solely social labels made users even more privacy preserving. In turn, using combined labels made it easier to identify users than when using four spatial labels. To use combined labels however, the compromise was made that a lot of time stamps should be discarded. This was first because the **dialogue** dataset was a combination of social datasets, which itself had few time stamps compared to the spatial datasets. Secondly, when combining the social and spatial labels, only observations which had common users and time stamps could be used, yielding less data even so. As the methods for obtaining mobility and social data were different, some users stood out in one dataset from the other, so a complete combination could not be made.

If one should generate such a combined dataset without compromising too much, missing data or labelled data could be inferred where none was obtained. For all datasets, another way of investigating missing data could indeed be to infer missing data as the most probable label for a user at given time bins based on either the probability of observing each label or from users who are the most similar to the user who needs data inferred. A simple similarity measure is the so-called *cosine similarity*

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}|| ||\mathbf{y}||} \tag{8.1}$$

where **x** and **y** are vectors. The similarity measure in (8.1) lies within the range $[-1; 1]$, and measures the angle between the vectors **x** and **y**. Therefore, a similarity of 1 means the vectors are exactly similar and a value of -1 means they are opposite [27, Ch. 2]. The similarity measure in (8.1) could therefore be used to find users who are the most similar, and then infer missing data for users based on other users they are the most similar to. This could however be hard if users have much missing data, but it could be a way to handle it.

As a way for finding how users' *typicality* affects their findability, likelihood estimations were performed for the **Home/Work** dataset, as seen in Section 4.4. The results indicated that using either the daily or weekly likelihoods did not matter, as they were highly correlated. However, results also suggested that the likelihood is negatively correlated with the findability, which contradicts the intuition. Intuitively, the more likely a user is to belong to a certain distribution, that is the more *average* or *typical* they are, the harder it should be to find them since they are more similar to other users. However, the results indicate that the higher the likelihood, the lower the 95th percentile and therefore the higher the findability.

These results cannot be clarified entirely, but it may have something to do with the way the likelihood distributions are constructed. As they are are constructed such that the distributions hold information on the percentages of being at home at each hour of the day or the week, being an average user means that one is mostly at home during the night hours and not as much in the daily hours.

As suggested by Figures 4.20 and 4.21, there is a large deviation in the middle of the day for the likelihood distributions, and since the log-likelihoods depend on the likelihoods for each hour of the day or week, this may influence the outcomes. Therefore, to get decent results that match the expectations, the likelihood distributions should be constructed in another way. Maybe more labels are needed in order to construct the functions, where one should find the likelihood of each label for all users. Then the likelihood for a single user could be a linear combination of the likelihoods for each label. Such an analysis has not been performed however, but could be worth looking into. Furthermore, it is not expected to be the method of finding the likelihood that is incorrect, since multiple methods have been used. Methods modelling the distributions to follow binomial and normal distributions have been used, and they showed similar results as the results seen in Section 4.4. However, as the raw distributions could not evidently be shown to follow either a binomial or normal distribution, the raw values of the histograms have been used instead. In any case, modelling users using likelihood functions is a good idea and would be an ideal method to investigate in further analysis.

The unicity results indicated that when using the combined labels rather than four spatial labels, users were more likely to be identified. As combined labels indeed does not mask the data such that privacy is more protected for users, acquiring such data can be difficult for attackers. If one should identify users in a combined dataset, the knowledge of both the whereabouts *and* the social status of a user is needed. For an attacker this may be hard to acquire, and indeed multiple datasets must be combined to achieve such a task of identifying users. Furthermore, the attacker must also know what the labels mean, that is if they are only presented as integers, letters or the like. Therefore, even though the masking yields easier identification when combining external knowledge, the attacker may have a hard time acquiring such knowledge which in turn yields the masking a good method of protecting privacy in *real life*.

8.2 Future work

To improve on the work in this thesis, a number of different methods could be applied in future works related to the analyses in this thesis. These methods will be discussed in this section.

8.2.1 Decomposition methods

A suggestion to alternate decomposition methods, where the Non-negative Matrix Factorization method was used on the **Home/Work** labels in Section 4.2 could be the popular *Principal Component Analysis* (PCA). PCA finds the directions (components) in space that account for the most variance where each component is orthogonal to the previous. This method can be used to find the dimensions that for example account for 90 % of the variance in the data. Such a method is valid as both a dimensionality reduction method and for visualization purposes, but also to find types of users, as these dimensions in space may represent different types of users. A drawback however is that PCA requires the components to be orthogonal, which might not always be the most optimal [27, App. B]. Another approach is therefore the so-called *Independent Component Analysis* (ICA), which does not assume the components to be orthogonal. ICA is mainly used as a *blind source separator*, meaning that it can for example separate two mixed sound signals into their single components by approaching the data as being non-Gaussian. ICA may therefore also be a candidate to inspect when trying to find types of users. Finally, a third approach could be to use the *Archetypal Analysis* (AA). In this approach all points can be described as a convex combination (linear combination where coefficients are non-negative and sum to 1) of the so-called *archetypes*. These archetypes can be seen as extreme points in the point cloud, which could be useful for finding types of users in either of the datasets used in this thesis [26, Ch. 14].

8.2.2 Models based on findability

As this thesis has provided a measure of findability for users in the form of the 95th percentile as defined in Definition 1, one could use this measure for either regression or classification purposes. It would be obvious to try simple regression models where the input is the labels and the output is the findability. Such data are of very high dimensions, so regression models that can handle such high dimensions should be chosen appropriately. An example of such a regression model could be the *elastic net* model, which performs both feature selection and can include more features than observations. Furthermore, it works well when features are correlated, which might be the case of users [26, Ch. 3].

When looking at classification methods, one could do something similar to the work done in Section 4.3, namely divide findabilities into groups and use those groups as classes. Other classification methods could be used on the same input, but it might be more interesting to find other features on which the findability of users depends. An example could be to find certain statistics of users, such as the percentage of being at home for different days of the week or the inter-event time between social events as input to the classifiers. Intuitively, finding characteristics that account for either low or high findability would be ideal for classifying properties and would be a good step forward based on the work in this thesis.

Chapter 9

Conclusion

This thesis has investigated how privacy is preserved in large collections of personal data in the special case of labelled data, where only raw values have been used in the literature. Data sets holding mobility records, social events and a combination of the two for users have been investigated to see the difference of how privacy is best kept. A thorough investigation of missing data was concluded, where different scenarios of handling missing data were established. A special case of findability was defined as the 95th percentile, which is a measure of how unique users are in contrast to other users. With this measurement, it was established that users are very different and that the scenarios discussed also were very different.

The most sensible scenario, which was Scenario 2, turned out to be the hardest scenario to identify users with in all datasets. In turn it was shown that Scenario 3 had the easiest identification of users, but this scenario ignores missing data completely, which actually may hold valuable information. Furthermore, using four spatial labels instead of two turned out to be less privacy-preserving, since an increase in resolution means an attacker hold more information on users and they are therefore easier to distinguish. Using four combined labels consisting of spatial and social labels also yielded easier identification rather than using four spatial labels. Also, using two labels for social events yielded the hardest identification of the datasets, promising the best way to keep knowledge of users hidden. Comparing these results to the results in the literature yielded that when aggregating over all users the same results were obtained, however the 95th percentiles yield more insight in users' behaviour, since it is possible to measure their findability based on their labels.

It was also discussed how users' diverseness affected the findability by using the Non-negative Matrix Factorization (NMF) together with classification methods to see that the NMF scores hold information on types of users. Likelihood estimations were also performed, however the results were counter-intuitive in regards to the expected results. However, it was concluded that either the likelihood in a daily or weekly resolution could be used, since they are highly correlated.

Finally, an analysis regarding the information on which time bins held more information than others was performed. This analysis, dubbed *temporal correlation analysis*, showed that indeed some time bins held more information than others when identifying users and that less information in general is present when considering multiple labels. Also, less temporal information is given for users who are easy to find, where more is given for users who are hard to find. Finally, when considering two labels the density, that is the probability of observing a 1, could be used to compare to the temporal correlations when averaging over users.

Appendix A

Pseudo-code for scenarios

Algorithms 1, 2 and 3 holds pseudo-code for Scenarios 1, 2 and 3 given in Section 3.1.

A.1 Algorithm 1

Algorithm 1 Identify uniqueness of user for Scenario 1

```
$N$ = number of users
$I$ = user of interest, denoted with a number between 0 and $N - 1$
S = $\{1, \dots, N\} \setminus \{I\}$, the set of users without $I$
$U$ = { }, the set of users to discard
$k$ = 1, number of points used
while S is not empty do
    $t$ = random time bin
    $L(I, t)$ = label of user $I$ at time $t$
    for all j ∈ S do
        $L(j, t)$ = label of user $j$ at time $t$
        if $L(I, t) \neq L(j, t)$ then
            move j from S to U
        end if
    end for
    save |S| at point $k$
    $k = k + 1$
end while
```

A.2 Algorithm 2

Algorithm 2 Identify uniqueness of user for Scenario 2

```
{Definitions are given in Algorithm 1}
while S is not empty do
    $t$ = random time bin where user $I$ does not have missing data
    $L(I,t)$ = label of user $I$ at time $t$
    for all j $\in$ S do
        $L(j,t)$ = label of user $j$ at time $t$
        if $L(j,t)$ is missing then
            go to the next user
        else if $L(I,t) \neq L(j,t)$ then
            move j from S to U
        end if
    end for
    save |S| at point $k$
    $k = k + 1$
end while
```

A.3 Algorithm 3

Algorithm 3 Identify uniqueness of user for Scenario 3

```
{Definitions are given in Algorithm 1}
while S is not empty do
    $t$ = random time bin where user $I$ does not have missing data
    $L(I,t)$ = label of user $I$ at time $t$
    for all j $\in$ S do
        $L(j,t)$ = label of user $j$ at time $t$
        if $L(j,t)$ is missing then
            move j from S to U
        else if $L(I,t) \neq L(j,t)$ then
            move j from S to U
        end if
    end for
    save |S| at point $k$
    $k = k + 1$
end while
```

Appendix B

Tables

B.1 Number of users in Home/Work dataset

Table B.1 shows how many users are left after preprocessing the **Home/Work** dataset given a minimum percentage of valid data.

Table B.1: *Number of users given different data qualities for the* ***Home/Work*** *dataset.*

Min. % valid data	Number of users
0	835
10	755
20	689
30	597
40	506
50	413
60	339
75	198
90	59

Appendix C

Sample code

C.1 Unicity function

The following Python script shows the main functions used when identifying users. The *uniqueness* function will find users for a given scenario and a given number of iterations. The results from the *uniqueness* function can then be used in the *plot95* function to plot a histogram of the 95th percentiles.

```
import numpy as np
import matplotlib.pyplot as plt
import pickle
import datetime
from collections import defaultdict
from collections import Counter
from collections import OrderedDict
from pylab import rcParams

def find(data,timestamps,users,labels,include_missing=True):
        """
        This function finds all users with given labels
        for all timestamps.

        Input:

        data                    :       Dataset, dictionary containing data.
                                                Keys are timestamps, values are labels
        timestamps              :       All timestamps considered in the dataset, list.
        users                   :       Users in the dataset, list.
        labels                  :       List of labels in dataset, list.
        include_missing :       Include missing values, i.e. label -1, boolean.

        Returns:

        s                       :       Users with label at all timestamps, dict.
        """
        if include_missing:
                labels.append(-1)
                s = {i: defaultdict(set) for i in labels}
                for u in users:
                        for t in timestamps:
                                s[data[u][t]][t].add(u)
        else:
                s = {i: defaultdict(set) for i in labels}
                for u in users:
                        for t in timestamps:
                                if data[u][t] == -1:
                                        continue
                                s[data[u][t]][t].add(u)

        return s

def uniqueness(data,it,timestamps,scenario,valid_values={},pmax=100,labels=[0,1]):
        """
        This function finds the number of users discarded at each iteration,
        as well as finding the fraction of users discarded at each iteration from
        the previous one.

        Input:

        data                    :       Dataset, dict of dicts containing data.
                                                Keys are users, values are dicts containing timestamps
                                                as keys and values as labels.
        it                              :       Number of iterations for each user, int.
        pmax                    :       Maximum number of points considered, int.
        timestamps              :       All time stamps considered in the dataset, list.
        scenario                :       Which scenario to consider, int 1,2,3.
                                                If there is no missing data, e.g. social
                                                datasets, use scenario = 1.
        labels                  :       List of labels in dataset.
        valid_values    :       Dict containing valid time bins for each user,
                                                keys are users, values are list of timestamps, e.g.
                                                {0: [t1,t4], 1: [t1,t2]}.

        Returns:

        size_dict       :       Number of users left after discarding at
                                                each iteration, dict.
        temporal_dict   :       Fraction of users discarded from the
                                                previous iteration, dict.
        """
        size_dict = {}
        temporal_dict = {"24": {}, "168": {}}
        users = list(data)
        if scenario not in [1,2,3]:
                raise ValueError("Invalid scenario chosen. Please choose between 1, 2 or 3.")

        print("Finding users...")
        if scenario in [1,2]:
                s = find(data=data,timestamps=timestamps,users=users,labels=labels)
        else: #scenario 3
                s = find(data=data,timestamps=timestamps,users=users,include_missing=False,labels=labels)
                #No need to keep missing data for other users
        no_users = 0
        print("Identifying users.")
        for single_user in users:
                print ("\rComplete: ", round(100*no_users/len(data),3), "%", end="") #prints progress
                no_users += 1
                temporal_dict["24"][single_user] = defaultdict(list)
                temporal_dict["168"][single_user] = defaultdict(list)
                ps = {}
                S_others = set(users) - {single_user}
                data_single = data[single_user]
                if scenario == 1:
                        bin_numbers = timestamps #use timestamps with missing data
                else:
                        bin_numbers = valid_values[single_user][:] #ignore missing data
                for i in range(it):
                        p = 0
                        p_list = []
                        S = set(S_others)
                        oldLen = len(S)
                        b = np.random.choice(bin_numbers,size=pmax,replace=False) #pmax random timestamps
                        dates = [datetime.datetime.fromtimestamp(i) for i in b] #timestamps in dates
                        values = {k: data_single[k] for k in b}
                        for j in b:
                                if len(S) == 0:
                                        break
                                t = dates[p]
                                val_user = values[j]
                                if scenario in [1,3]:
                                        S = S & s[val_user][j] #keep users with same value
                                else: #scenario 2
                                        S = S & (s[val_user][j] | s[-1][j])  #keep users with same value or
                                              if they have missing data in j
                                p_list.append(len(S))
                                temporal_dict["24"][single_user][t.hour].append(len(S)/oldLen)
                                temporal_dict["168"][single_user][t.weekday()*24 + t.hour]\
                                        .append(len(S)/oldLen)
                                oldLen = len(S)
                                p += 1
                        ps[i] = p_list
                size_dict[single_user] = ps
        print ("\r100%", end="\n")

        return size_dict, temporal_dict

def save_pickle(file,object):
        """Function used to save object to a binary pickle file.
        Input:

        file                    :       Name of file to write, string.
        object                  :       Object to save.
        """
        try:
                f = open(file,"wb")
        except:
                raise FileNotFoundError("File not found.")
        pickle.dump(object,f,protocol=pickle.HIGHEST_PROTOCOL)
        f.close()
        return

def load_pickle(file):
        """Function used to save object to a binary pickle file.
        Input:

        file                            :       Name of file to read, string.

        Returns:

        object                          :       Object to load.
        """
        try:
                f = open(file,"rb")
        except:
                raise FileNotFoundError("File not found.")
        result = pickle.load(f)
        f.close()

        return result


def percentile95(results_all):
        """
        Calculate 95th percentile for all users
        based on the results from "uniqueness".

        Input:

        results_all                     :       Results from "uniqueness", dict.

        Returns:

        percentile95            :       95th percentile for all users, dict.
        """
        max_points = defaultdict(list)
        for user in results_all:
                for iteration in results_all[user]:
                        max_points[user].append(len(results_all[user][iteration]))

        percentile95 = OrderedDict((i,np.percentile(max_points[i],95)) for i in max_points)

        return percentile95

def plot95(data,max_bound=100,title="",save_title="",save=True):
        """
        Plots the 95th percentile of all users in a histogram.

        Input:

        data                            :       Results obtained from "uniqueness", dict
        max_bound                       :       Max. number of points in plot, int.
        title                           :       Title of the figure, string.
        save_title                      :       Destination of the file if "save==True", string.
        save                            :       Whether to save the plot or not, boolean.
        """
        p95 = percentile95(data)
        c = Counter(p95.values())
        c2 = defaultdict(int)
        for i in c:
                c2[round(i)] += c[i]

        font = {'size'  : 22} #font size on plot
        plt.rc('font', **font)
        rcParams['figure.figsize'] = 12,6 #figure size
        plt.bar(list(c2.keys()),list(c2.values()),width=1,align="center")
        plt.title(title)
        plt.xlabel("95th percentile")
        plt.ylabel("Number of users")
        plt.xlim([-1,max_bound+1])
        plt.xticks(np.arange(0,max_bound+1,10))
        if save:
                plt.savefig(save_title)

        plt.show()
        return
```

Bibliography

[1] CloudTweaks. Surprising facts and stats aboout the big data industry. https://cloudtweaks.com/2015/03/surprising-facts-and-stats-about-the-big-data-industry/, March 2015. Retrieved May 10th 2017.

[2] Harry E. Pence. Will big data mean the end of privacy? *Journal of Educational Technology*, 44(2):253–267, 2015.

[3] Fengli Xu, Zhen Tu, Yong Li, Pengyu Zhang, Xiaoming Fu, and Depeng Jin. Trajectory recovery from ash: User privacy is NOT preserved in aggregated mobility data. *CoRR*, abs/1702.06270, 2017.

[4] Yves-Alexandre de Montjoye, Erez Shmueli, Samuel S. Wang, and Alex "Sandy" Pentland. openpds: Protecting the privacy of metadata through safeanswers. *PLOS ONE*, 9(7):1–9, 07 2014.

[5] Guy Zyskind, Oz Nathan, and Alex "Sandy" Pentland. Decentralizing privacy: Using blockchain to protect personal data. In *Security and Privacy Workshops (SPW), 2015 IEEE*, pages 180–184. IEEE, 2015.

[6] United Nations General Assambly. Universal declaration of human rights. http://www.un.org/en/universal-declaration-human-rights/index.html. Retrieved May 5th 2017.

[7] CNN. Comparisons between '1984' and today. http://edition.cnn.com/2013/08/03/opinion/beale-1984-now/. Retrieved May 5th 2017.

[8] Pew Research Center. The state of privacy in post-Snowden America. http://www.pewresearch.org/fact-tank/2016/09/21/the-state-of-privacy-in-america/, September 2016. Retrieved May 10th 2017.

[9] Kenneth Einar Himma and Herman T. Tavani. *The Handbook of Information and Computer Ethics*. John Wiley & Sons, Inc., 2008. Electronic edition in PDF format retrieved at http://onlinelibrary.wiley.com/ on May 1st 2017 through the DTU findit search engine.

[10] amiunique. Browser fingerprint collection website. http://amiunique.org. Retrieved May 2nd 2017.

[11] Peter Eckersley. How unique is your web browser? In *Privacy Enhancing Technologies, 10th International Symposium, PETS 2010, Berlin*, pages 1–18, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[12] Jagdish Prasad Achara, Gergely Acs, and Claude Castelluccia. On the unicity of smartphone applications. In *Proceedings of the 14th ACM Workshop on Privacy in the Electronic Society*, pages 27–36. ACM, 2015.

[13] Business Insider. Google knows where you have been. http://www.businessinsider.com/how-to-see-where-google-knows-ive-been-2015-2?r=US&IR=T&IR=T. Retrieved May 5th 2017.

[14] Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Sci. Rep*, 3(1376), 2013.

[15] Yves-Alexandre de Montjoye, Laura Radaelli, Vivek Kumar Singh, and Alex "Sandy" Pentland. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 347(6221):536–539, 2015.

[16] Rui Martins. Privacy-respecting features in large collections of personal data. Master's thesis, Technical University of Denmark (DTU), June 2016.

[17] Continuum Analytics. Python programming language distribution. https://www.continuum.io/downloads. Retrieved January 10th 2017.

[18] Sensible DTU. Official website for the Sensible DTU study. https://www.sensible.dtu.dk/. Retrieved May 7th 2017.

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

[20] Piotr Sapiezynski, Arkadiusz Stopczynski, Radu Gatej, and Sune Lehmann. Tracking human mobility using wifi signals. *PloS one*, 10(7):e0130824, 2015.

[21] Vedran Sekara, Arkadiusz Stopczynski, and Sune Lehmann. Fundamental structures of dynamic social networks. *Proceedings of the national academy of sciences*, 113(36):9977–9982, 2016.

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

[23] Sune Lehmann Jørgensen. List of Sensible DTU publications. https://sunelehmann.com/sensibledtu-publications/. Retrieved May 8th 2017.

[24] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[25] Sci kit Learn. Nmf implementation by sci-kit learn. http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html. Retrieved February 14th 2017.

[26] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, 2 edition, 2009.

[27] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Addison-Wesley, international edition edition, 2006.

[28] Sci kit Learn. Cross-validation discussion by sci-kit learn. http://scikit-learn.org/stable/modules/cross_validation.html. Retrieved March 6th 2017.

[29] Sci kit Learn. Knn implementation by sci-kit learn. http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html. Retrieved March 6th 2017.

[30] Sci kit Learn. Random forest implementation by sci-kit learn. http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. Retrieved March 6th 2017.

[31] Scipy. Spearman correlation function by Scipy. https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.spearmanr.html. Retrieved March 8th 2017.