Predictability of Human Mobility from Highly Granular Location Data

Rafaela-Ioana Voiculescu

Kongens Lyngby 2014

Technical University of Denmark
DTU Compute - Department of Applied Mathematics and Computer Science
Richard Petersens Plads, Building 324,
DK-2800 Kgs. Lyngby, Denmark
Phone: +45 45 25 30 31
compute@compute.dtu.dk
www.compute.dtu.dk

Summary - English

The understanding of human mobility patterns constitutes a focus point in the science world today and has been the main topic of numerous studies. The data that has been used in order to analyze the mobility patterns varies from GPS data, to data collected by telecommunication towers and even to information extracted from the dispersion of bank notes. In this paper we study the inferring of mobility patterns from Wifi data. We analyze methods that can be used for identifying stop locations with the help of Wifi data, we compare our results with stop locations identified using GPS data and we explore the degree of predictability of human travel trajectories.

Summary - Danish

Forståelsen af menneskers bevægelsesmønstre er et aktivt forskningsfelt og omdrejningspunkt for talrige studier. Datamaterialet bag disse analyser spænder bredt, fra GPS-data til data indsamlet via telekommunikationsmaster, og helt til bevægelsesmønstre udledt på baggrund af diffusion af pengesedler. I denne afhandling ser vi nærmere på hvordan bevægelsesmønstre kan udledes fra WiFi data. Vi analyserer metoder, som på baggrund WiFi data, kan bruges til at identificere "stop locations" – steder hvor en person opholder sig i længere tid – og sammenholder disse resultater med "stop locations" identificeret ved hjælp af GPS data. Endelig undersøger vi i hvilken grad menneskers bevægelsesmønstre kan forudsiges.

Preface

This thesis was prepared at the Department of Applied Mathematics and Computer Science at the Technical University of Denmark in fulfilment of the requirements for acquiring an M.Sc. in Computer Science and Engineering.

The thesis describes the steps taken in order to explore methods which can be used in order to determine stop locations as well as the predictability of human mobility based on Wifi data.

Lyngby, 01-August-2014

Rafaela-Ioana Voiculescu

Acknowledgements

I would like to thank my supervisors, Sune Lehmann and Jakob Eg Larsen, for their continuous guidance and help throughout the course of the work for this thesis.

I would also like to thank Andrea Cuttone, Piotr Sapieżyński, and David Kofoed Wind for their feedback, suggestions and assistance with technical issues.

Finally, I would like to thank my family and friends for their support and encouragement.

Contents

Chapter 1

Introduction

The United Nations (UN) Department of Economic and Social Affairs' Population Division is in charge of preparing once every two years an estimation of what we are to expect the growth rate for the world population to be in the following years. Occasionally, world population projections over a longer period of time are created as well. In 2004, the UN has made predictions about the trends in population growth up until 2300 [UNW]. The predictions show that the population will increase to reach a peak of 9.22 billion by the time of 2075 and then it will decrease slowly to 8.97 billion by 2300. These estimations bring up different aspects regarding our quality of life which is a topic that will only increase in importance with the population growth. For example, we need to think about more efficient ways in which we can develop the urban regions or our transportation infrastructure in order to ensure that space is used in a responsible and optimal manner. Also, it is important to understand how epidemics spread and how they can be contained in order to avoid devastating pandemics from tacking place.

The rapid evolution of technology has equipped us with tools that can be used in order to gather large amounts of information about human behaviour and mobility patterns. This information can be employed by scientists in research and study programs in order to obtain facts and results that can be used to solve possible problems which are sure to appear due to the rise in population that seems to be facing us. Current studies about the predictability of human

mobility and our travel trajectories uncovered remarkable findings which suggest that we might be less spontaneous when it comes to choosing our destinations than we might think we are. These results can prove to be of tremendous help in the making of decisions about transportation infrastructure and they can even give us an insight on how diseases are spreading from region to region.

Due to the importance of the results that can emerge from studying the predictability of human mobility, we have conducted a study that is focused on this topic, mainly the inferring of mobility patterns from Wifi data which has been gathered from volunteers. In order to understand what defines a location we analyze different methods that can be used for extracting stop locations from gathered Wifi data. We discuss and propose a solution for determining if two identified locations represent in fact the same geographical stop location. This helps us construct a long term image of the travel trajectories associated to the users who are providing the data. We compare the results we have, our stop locations obtained from the Wifi data, to results generated using GPS data and we explore the degree of predictability of human travel trajectories based on the image we are able to create about the mobility patterns over a longer period of time.

The detailed explanation of all steps made during our work as well as the results and observations we have come across are structured into eight different chapters. Chapter 2 (Related work) presents previous findings and studies conducted on the topic. Chapter 3 (Prerequisites and tools) presents the elements that have contributed to making this work possible: data gathered with the help of volunteers, implementation, visualization and data analysis tools etc. Chapter 4 (Data processing) presents the data that has been used during our work as well as the way in which we have eliminated interferences and noise from the received data. Chapter 5 (Extracting locations from Wifi data) presents all the steps that have been taken from the data analysis up to the analysis of the different algorithms that we have experimented with in order to extract the locations from the Wifi data. Chapter 6 (Location matching) presents the techniques we have considered in order to estimate if two given locations that have been discovered by using a location extraction algorithm can be catalogued as being the same location based on defining characteristics. Chapter 7 (Entropy and predictability) presents the calculations made in order to determine the different entropy values and predictability that can be attributed to the users based on the location information that emerged from their data as well as what these values represent. Chapter 8 (An evaluation of Wifi positioning accuracy) presents the comparison made between the results we have obtained from the Wifi data and the stop locations which can be extracted from the analogue GPS data. Chapter 9 (Discussion and future work) presents a summary of the results and observations which have emerged during our work as well as proposed topics and directions for further work on the present subject.

Chapter 2

Related work

There is a high interest and a huge amount of work the scientific community dedicates to understanding the patterns of human mobility. The knowledge we can gain from the results of this work has the potential to benefit a wide variety of industries from the modeling and maintenance of the transportation infrastructure, to the medical industry where we can use this knowledge in trying to prevent the spreading of epidemics [DB08].

Before starting the work on any subject we first need to understand what has been done previously, what were the results and what have been the suggestions for future research. The present chapter is dedicated to presenting previous studies and the findings they offer us. The studies have been divided into 6 sections. The first two sections focus on previous studies which provide knowledge about mobility patterns and behaviour that can be attributed mostly to communities and not individuals. The following three sections present studies that focus on individual mobility patterns. The last section presents the results of studies that have focused on analysing and calculating the predictability of human travel trajectories.

2.1 Mobility patters uncovered by the dissipation on bank notes

Brockmann et. al.[DB06] have analyzed the human movement based on the way bank notes were dispersed through the United States (excluding Alaska and Hawaii). Their study shows that a relatively small percentage of bank notes (23.6%) traveled for more than 800 km, while a fraction of 19.1% did not traveled for more than 50 km even after a year of being observed. The possible explanations the authors have given for these findings are that, in general, people would be less inclined to leave the areas of the large cities or the places they usually conduct their lives.

The problem identified with this approach for tracking individuals is that the bank notes exchange hands and the behaviour which is identified by the way they circulate can't be attributed to a single individual, but rather to different ones that at any moment have had the bank note in their possession. Despite this, the result have a high scientific value as they do identify patterns in human travel behaviours in general.

2.2 Eigenbehaviours

Eagle et. al. analyze data of individuals and communities with the purpose of trying to predict and cluster the daily habits and behaviour of people [NE09]. They consider that the behaviour of one person throughout a day can be close to a sum of their primary eigenbahaviours throughout that day. The results of the study have shown that when having a weighted sum calculated for the first half of a day, the behaviour of the same person throughout the remaining of the day can actually be approximated with 79% accuracy.

The results have applicability in more fields, as they allow us to consider the possibility of clustering people into various communities based on the similarity of their behaviours. It goes even further, as the findings show that this enables the possibility of calculating similarity for groups as well and thus permitting a classification that, according to the experiment, can be 96% accurate for determining affiliations in the social network of a particular population.

As a last observation in the paper by Eagle et. al. it is stated that eigenbehaviours can be used in order to identify the possible friendship ties between people. The observations in this paper have been done based on the Reality

Mining data set that tracked the behavior for 100 individuals at MIT for the duration of one year.

2.3 Mobility patterns of mobile phone users

Barabasi et. al. have conducted a study [MCG08] that deals with observing the trajectories of over 100000 mobile phone users with annonymized identities. The study was conducted in order to see if there are any patterns in our mobility habits. Among the things that have constituted subject for testing was the return probability of individuals in a previous place. The study shows there is, in general, a peak in the return probability after 24, 48 or 72 since people have left a particular location. This shows that we tend to visit locations periodically. This can be explained by our needs of going to places such as work, school, grocery shops near our home etc.

The authors have also ranked the locations the mobile phone users frequented based on the number of times they have been spotted nearby. The results for this experiment have shown that the probability of finding someone near a location that is ranked for them with a level L can be estimated with $1/L$. Another interesting finding that is mentioned in the paper is that, in general, people seem to be spending the majority of their time in just a few locations, while dividing the remaining time just between a limited number of locations that varies for the subjects from as low as 5 to around 50.

The results of this study are a major indicator that individuals display a high level of regularity and that we have a tendency to spend most of our times in places that are familiar to us, or that require us to visit them regularly (e.g. home, work).

2.4 Human movement recorded through real traces

Studies as the ones with the travel of bank notes or the recorded location of mobile users through telephone cannot be very exact and usually cannot reflect the real traces for the people taking part in them. They provide a very useful estimation, however with the technology that we have access to nowadays, we are able to record mobile phone users' real traces either through GPS or Wifi. The data that can be acquired through these means allows us to conduct studies that can take into consideration an even better approximation of the real location of individuals.

In the paper by Kim et. al. [MK06], the authors present us with a method in which the locations of users can be estimated based on the WiFi signals that their devices register. The experiment is conducted considering the data for a duration of 13 months. The user traces that have been used consist of the trace data from Dartmouth College. The mobility traces are defined as the lists of access points that are associated to a user's devices at a given timestamp.

The mobility traces allowed the authors to extract the tracks (locations) of the users. They have explored three methods in which the location can be extracted from the data. The first approach presumed the calculation of the center (intersection of medians) of the triangle defined by the past three access point associations of the mobile device of the user. This approach has a downside since the devices do not necessarily change the associations in a periodic manner. This lead to the second approach which consisted in considering a time window after which the associations needed to be updated in case new associations have appeared during that time. The third and last approach explored the use of Kalman filters [Kal60].

In order tp validate the path extractors the authors have compared the results with GPS data. This validation has proved that the type of device used for collecting the data can have a significant importance in how accurate the results are as it seems that some devices can be more aggressive in updating the associations with access points while others try to stay associated with the same access points as long as possible before switching to new ones. This leads to problems as different distances between users and access points considered by different devices and as such it affects the estimated paths. The best estimations have been given in this experiment by the approach that used the Kalman filters, however both the other two approaches have provided fairly good estimations as well.

Another paper which explores the travel patterns from real data is the one written by Azevedo et. al. [TSA09]. The authors propose another approach for analyzing the mobility of people. They take into consideration the following movement components: velocity, acceleration, direction angle change and the pause time and they are using the GPS data in order to estimate the locations of individuals. The experiment takes place in a park in Rio de Janeiro and is done based on the data received from approximately 120 volunteers. The results have shown that people seem to have in general smooth trajectories without abrupt changes.

2.5 Mobility patterns in massive multiplayer online games

Sinatra et. al. have studied the way in which users of a massive multiplayer online game behave inside the virtual universe provided by the mentioned game [RS14]. It has been established that the massive multiplayer games provide people with a virtual reality where they can interact with others through their characters and can, in fact, form groups and, as such, display both individual as well as collective behaviour actions that can translate to the non-virtual world [Bal03].

This study gives an interesting insight into the habits and actions of the characters which are controlled by the players. Among the things the authors have analyzed are the predictability of the characters, the entropy generated by the mobility of the characters in the virtual universe and general strategies or patterns that could be observed.

The game the authors have been using for the study is called Pardus [Par]. This game is quite complex, as it allows the manifestation of normal real-life activities such as the creation of alliances or friendships, communication between the players, economic related action, or even actions which have a negative connotation such as attack of another user, removal of a friendship link etc. The universe of the game consists in hundreds of nodes which represent cities or sectors in the game. These virtual cities are tied to each other through links which mark the possibility for the users to move their characters from one place to another.

By analyzing the why in which characters have interacted through the years, the authors have observed that the mobility of the characters through the universe is highly predictable, as users in general will seem to be choosing a next location in a random manner in just about 10% of the cases.

2.6 Entropy and predictability

One step further from understanding the way we travel from place to place is to predict our future locations based on a previous knowledge of our past patterns. There has been an extensive study done in this area of the scientific playground as well and the results which have emerged up until now are remarkable.

In the paper by Song et. al. [CS10], the authors take up the challenge of

studying how predictable people can be. They analyze the mobility patterns of mobile phone users and calculate the entropy of these users. The locations are defined by the telephone towers the users are encountering at hourly intervals while the trajectory of the user is given by the ordered sequence of these towers. The real entropy of each user i is calculated as $\Sigma_{T_i' \subset T_i} P(T_i') log_2(P(T_i'))$, where $P(T_i')$ represents the probability of encountering a time-ordered subsequence T_i' in the sequence of hourly encountered telephone towers T_i.

The results for this particular study show that, for the considered users, the uncertainty of where they could be at a certain moment, based on the real entropy calculated for them would be very low as they would most probably be in one of two locations.

The authors also take a look into the maximum predictability which can be expected for a user. Their results show that, with the right algorithm, a user's future location can be predicted with between $80 - 93\%$ accuracy. This shows that we are less spontaneous than we might think and that our mobility patterns are, in most cases, rooted into a very well established routine.

There have been numerous other methods or experiments conducted in order to analyze or to forecast human mobility patterns. A large number of the existing studies focus on identifying stop locations by analyzing GPS data. There are different types of algorithms which can be used to separate the GPS data into stop locations, for example distance grouping [CLL14], speed based DBSCAN clustering [PBKA08], the usage of GPS signal loss [CCJ10], the k-means clustering [AS03], etc. Some other studies include the Markov chain models [Ros09] [GL96], the neural networks [SCL03] or the Bayesian networks [AS07], finite automaton [JP04] or the study of human walks traces [LHK+09] [RSH+08]. Most of the studies support the idea that people's actions and travel behavior is indeed far from being random and thus the science world needs to dedicate further effort and time in order to continue the work done for understanding the way in which mobility patterns emerge so that we can use this knowledge to improve our quality of life and the world we live in.

Chapter 3

Prerequisites and tools

In order to research the way in which people travel we need to have access to a database of information that can be used for this purpose. As it was mentioned in Chapter 2, scientists have been trying in numerous way to identify and work with location information. The current chapter presents our decision of using Wifi data for observing mobility patterns, the way in which we have accessed the data for the study as well as some data manipulation and visualization tools that have come to be of use to us throughout our work.

During our study, we have dedicated our time in working with information about the access points that were visible to the users' mobile phones throughout their day. This has allowed us to implement and analyze different ways in which locations can be extracted from such information.

3.1 SensibleDTU

The data we are using is part of a large-scale study that aims to make observations based on the actions of volunteering students - the Copenhagen Network Study [SSS+14]. The main aim of this project is to offer an extensible framework for different studies that can help us have a better understanding of the human

nature and how personal data can be used to analyse individual behaviour in order to promote self-awareness and positive behaviour changes. The deployments for this study from 2012 and 2013 are based at the Technical University of Denmark and are named SensibleDTU.

A factor that represents a definite strength of this project is that it allows the phase of data collection to co-exist with various analysis phases. The platform allows the conducting of controlled studies which can be distributed to participants with the help of the smartphone software. Due to the way in which the platform is designed, the participants can be divided in different groups and these groups can be exposed to different stimuli in order to allow the analysis of results gathered for a multitude of possible experiments. Another benefit of being able to work with the provided data as soon as it is received is that data can be monitored without unnecessary delay. By doing this, the quality of the date can be evaluated for both the level of an individual user as well as for the whole dataset. The results of a qualitative evaluation allow the researches to understand if the collected data in its form is sufficient for answering the various research questions. The real-time processing of the data has the benefit that it allows the researches to create applications and services designed for the participants. These services can be used for receiving feedback and thus improve the understanding of what can be improved in order to maximize the usefulness of how the data is used in the ongoing studies, but they can also be used for offering the participants the chance to access the results obtained from their own data and thus helping them understand themselves better.

The students that consented to being volunteers for SensibleDTU have received smartphones that are able to track different aspects of their lives and through which they can interact with the system. The big number of volunteers [1] has allowed the gathering of a considerable amount of data regarding the mobile phone users' behaviour.

The data is collected through a variety of methods. Participants in the various studies can receive questionnaires which can be focused on getting information about their socio-economic background, habits or even psychological traces [2]. Sensor data is collected throughout the days at regular intervals as well as Wifi data. SensibleDTU offers the participants the option to authorize the collection of data from Facebook. This type of data can be used in order to create friendship graphs and to follow the interactions between various participants in

[1] Up to the moment at which the present paper has been written, there have been deployed over 1000 smartphones to students who wanted to take part in the study.

[2] For example, a survey was presented to the participants in 2012 consisting of over 90 questions. In 2013 an addition of over 300 questions were asked per participant. The questions targeted different aspects from working habits and various socio-econimic factors to Big Five Inventory measuring personality traits [JS99] and self-esteem.

the studies. Qualitative data is collected in order to gather feedback from the participants as well as understanding what can keep their interest in the project at a high level.

Since the majority of the collected information about the students is sensitive [AS14], keeping the data secure is and has been a top priority from the beginning of the experiment. The data is annonymized and stored securely. Another positive aspect is the fact that the students that are part of the experiments have access to tools that allow them to see what data they are sharing, what it is done with this data and which allow them to control how much they want to share. The combination between the measures that are taken in order to keep the data annonymized as well as the fact that the users have control and understanding of their own data ensures a good security level.

3.2 Using Wifi data for defining locations

The main aim of the present project is to explore the human mobility and the predictability of our traveling destinations based on the Wifi data gathered for a selection of participants from the SensibleDTU study.

The market of smart hand held devices is constantly and rapidly growing. This has opened a new market for applications which take the location of the user into consideration for offering him or her personalized services [ASA10]. In order for information about locations to be offered to such applications, the need for a system that can identify these locations at a low cost and in a sufficiently accurate manner has developed.

"Wireless Positioning Systems (WPS) provide a position estimate based on the radio signals received at a given location (measurement), and a known radio map of the environment."[AGGP09] The possibility of determining locations based on Wifi data is based on the fact that each access point that can be found inside a network has an unique id attributed to it (BSSID) and that each location receives a different signal strength from an access point given the distance in between them. Thus, the received signal strength (RSS) can also prove to be an important factor in the definition of a location based on Wifi data and various algorithm for determining position based on Wifi have already been designed and used [AGGP09].

Studies show that Wifi positioning can be acceptably good, as it can have an accuracy of $1 - 4$ m in an indoor environment and between $10 - 40$ m for an outdoor environment, depending on the number of access points which define

the area as well as the presence of interferences that can affect the accuracy for the positioning [CCLK05] [MR07]. Considering these numbers, the increase in the usage of smartphones and the cost wise advantages presented by Wifi positioning systems, the studies on human mobility include and should keep including the data that can be acquired from Wifi networks.

3.3 Implementation tools

Before starting the work on the present research, we have carefully taken into consideration possible tools that can be useful in our work.

The scripts that are used for analyzing, transforming and working with the data are developed in Python. The reasons behind using Python instead of any other programming language are numerous. Python is elegant and simple to use, it allows fast development and the code can be easily adapted and reused. Due to its high scalability, it is the perfect choice for both large and small projects, being easily extensible at the same time. Another very important reason for using Python is that there is a large number of libraries that can be used with it and that allow the visualization or handling of big data [3].

The matplotlib library [Mpl] is a very useful tool when dealing with visualizations which can be made directly from Python scripts and that can be used for data analysis. It can be used in order to generate plots, histograms, scatterplots, bar charts and many additional types of figures in an easy way.

An additional tool that has been used for the present project is Gephi [Gep]. Gephi is a platform that allows the exploration and handling of various networks and graphs. Further information on how this tool has proved helpful can be found in Chapter 5.

Scikit-learn [SL] provides ready, simple to handle and effective tools that can be used for data mining and analysis. The provided tools include regression algorithms, implementations for various models (e.g. Hidden Markov Models, further details in 5.3.4), pre-processing for various data, clustering algorithms (e.g. k-means, further details in 5.3.3) and many other equally useful mechanisms and implementations.

Pandas [Pan] is a powerful and easy-to-use Python library that allows the work with data structures through offering various data analysis tools. Among the tools provided by this library we can find the intelligent data alignment, tools

[3] Examples of libraries and packages used: numpy, pickle, datetime, sympy etc.

for reading and writing data of various formats, merging and joining of data sets, grouping and reshaping of data and many others.

Chapter 4

Data processing

Data processing is an important part of any project that relays on big amounts of data. The present project uses data that has been selected from the database of the SensibleDTU experiment. The data is fully annonymized and the users that have been a part of the study have been chosen randomly from the database. The current chapter presents the structure of the various input data that we have used throughout our work, as well as the way in which we have conducted the data clean up in order to remove noise and interferences.

4.1 Statistics

We use the data collected from 131 users from the SensibleDTU database. The students that have been selected for the present study have had data collected for a period of almost a year. [1]

The application that is installed on the smartphones of the students who are part of the experiment is configured to scan periodically (around every 15 seconds) for Wifi networks, however, it is also set to record the scans which are triggered by any of the other applications that are present on the mobile phone.

[1] Data has been collected from 2012 up to September 2013.

4.2 Wifi and GPS data

For the present study we are not using all the fields that are accessible from the database of collected information. The aim of the study is to analyze the predictability and patterns in the human mobility and as such we need information that can help us identify the locations of the users that are part of the study. For this we are accessing fields related to the *Wifi information* associated to the selected group of users. The results regarding the users' locations over time are afterwards compared with locations extracted from *GPS data* and as such we are accessing this information from the database as well.

For working with the Wifi information that is available in order to identify user locations, we extract from the database the fields that can be seen in Tab. 4.1.

user	timestamp	ssid	bssid	rssi	context
1	1349185621	1	1	-75	0
1	1349185685	4	4	-86	0
1	1349185700	5	5	-84	0

Table 4.1: This table shows a few examples of possible Wifi data recorded from users

A short explanation for each of the fields can be found below:

- The user (first) field gives us information about what user we are currently observing. The real identities of the users are concealed and replaced by an ID which is unique for each of them.
- The timestamp (second) field gives us information about the moment of time at which the scan occurred and for which the information is gathered. The time format is Unix timestamp. [2] This timestamp can be easily manipulated and converted to any other timestamp format in Python by using the datetime module that can be found in the Python Standard Library [PSL].
- The SSID (third) field stands for Service Set Identifier and it represents the unique ID that can be used in order to identify the wireless networks. This identifier is responsible for the correct sending of data when multiple wireless networks overlap.
- The BSSID (forth) field stands for Basic Service Set identifier and it represents the MAC address of a wireless access point.

[2] The Unix time stamp represents a way in which time can be tracked as the total number of seconds starting from January 1^{st}, 1970 at UTC and a particular date and time.

- The RSSI (fifth) field stands for Received Signal Strength Indication and it represents the strength for a signal picked up by the mobile phone from an access point. The RSSI values in our case are registered as the real signal strength recorded in dBm and are therefore negative values. As such, the signal is stronger when the value recorded for it is closer to 0.
- The context (sixth) field is based on the SSID and it translates to the possibilities presented in Tab. 4.2

context	translation
0	unknown
1	AndroidAP
2	eduroam
3	dtu
4	device
5	eksamen
6	iPhone
7	Bedrebustur (wifi on bus)
8	CommuteNet (wifi on train)

Table 4.2: This table shows the possible contexts for the retrieved Wifi information from the students

The GPS data we are using in order to compare our identified Wifi locations has the fields which can be observed in Tab. 4.3.

user	timestamp	latitude	longitude	accuracy	provider
1	1349196008	55.6752334954	12.3736925237	10.0	gps
3	1349196332	55.7284276932	12.5028730743	5.0	gps
93	1350152289	55.4494251078	12.1857996751	10.0	gps

Table 4.3: This table shows a few examples of possible GPS data recorded from users

A short explanation of each of the fields is as follows:

- The user (first) field, similar to the Wifi data, gives an unique annonymized identifier for the user to which the data corresponds.
- The latitude (second) field gives the value of the latitude for the place at which the scan occurred
- The longitude (third) field gives the value of the longitude at which the scan occurred

- The accuracy (forth) field gives information about how accurate the position identified in the scan is (it is measured in meters)
- The provider (fifth) field informs about the entity which provided the data for the scan. The providers can be GPS, Wifi or celltower.

4.3 Interferences in Wifi networks

Nowadays Wifi networks are used for a multitude of activities from web browsing to video viewing and even to voice or text communication between people all over the world. As the usage of this technology is expanding so does the need for an even more reliable Wifi connection service. The current issue with the Wifi networks is that they are using the IEEE 802.11 protocol [WLP] that uses the 2.4 GHz Industrial, Scientific and Medical Radio Frequency band [Fli03]. This band is, however, unlicensed which means that various devices (Wifi and non-Wifi alike) can use it. This leads to the apparition of interferences.

The results of the experiment conducted by Mahanti et. al. [MCWA10] show that a variety of factors can affect the Wifi networks transmission and signal strengths. For example, microwave ovens, analog wireless video cameras, analog cordless phones and wireless jammers can have a severe impact on the Wifi operations.

However, the issue that causes the most problems in our data set is the existence of signals that come from access points which can be observed for just a very short period of time as they or the user quickly move by, or that are sufficiently far away from the device and as such their signal level is very low and they can periodically be missing from the scanned access points even when the users find themselves in the same location [FBSW08].

4.3.1 Assumptions about noise and initial data cleaning

Before starting to eliminate the noise in our Wifi data, we have made a few assumptions on what is to be considered noise in the data for the present study. The assumptions are as follows:

- Data received from access points that are part of bus or train Wifi networks are to be ignored (meaning entries that have the context number set to 7 or 8). This assumption was made as it would be hard to determine the

characteristics of a given location considering the access points present in buses or trains. For example, a person can take different buses which have a common portion of a route, yet the access points identified by the phone would be completely different and thus the locations would be impossible to be matched based only on this information. Also, not to mention that the bus or train would be moving.

- Data received from hot spots created from Android or iPhone devices (entries that have the context number set to 1 or 6) can also be ignored. These access points are most probably mobile and will not be present in the same locations. This means that they are not reliable when defining locations based on the Wifi networks visible to the mobile phones.
- The signal strength of the registered access points can give information about the distance between the device and the access points and as such it can be a factor in determining what access points need to be taken into consideration when computing the locations. The paper by Zhang et. al. [ZF12] presents the POLARIS system that aims to deal with localization based on Wifi and it also deals with eliminating noise or disturbances from the data. They consider that any signal that has the signal strength indication outside the range of -60 to -99 dBm can be catalogued as signal disturbances. However, during our data analysis we have observed that the devices can register signals that have a RSSI value above -60 (which means that the signal is more powerful) and as such, for our data we consider just the lower bound of -99 dBm as a limit for noise. The data registered for access points that have an RSSI value below this one are ignored in order to helps ensure that only the access points which are acceptably close to the device are taken into consideration when trying to determine the location of the user.

4.3.2 Data cleaning

Data cleaning is important as inconsistent or incorrect data might lead to inaccurate conclusions and observations. Considering this, the noise elimination in Wifi data is of high importance. Keeping in mind the previously made assumptions (Section 4.3.1), we have eliminated the entries that did not respect the previously mentioned criteria.

During our work with the data, however, we have observed that there are other cases in which additional problems can appear. These situations have been encountered when dealing with the extraction of the mobile users' locations from the information we have regarding the associations made between their phones and various access points. Some of the algorithms used for computing

the locations are very time consuming and as such the presence of unnecessary data can burden even further the analysis causing an exponential increase in the execution time.

The situations in which we can struggle with data that does not give any additional information for the identification of locations and that are not necessarily solved by the noise elimination done based on the assumptions presented in the previous sections are caused by the existence of what we will name *isolated observable access points*. We define as isolated observable access points the access points that are visible to the mobile devices for a very short period of time after which they stop being visible for a long period of time. The reason behind the access point not being visible for longer periods of time can be varied, for example: defective access point, the distance between the access point and the user is increasing very fast in the short period of time between scans etc.

Fig. 4.1 illustrates a possible case in which these access points can cause problems rather than help. As we can see there are 7 access points that have appeared in the mobile phone scans over a period of one day. Let us consider that access point AP1 only appears during two consecutive scans, however, it will be taken into account when computing the locations that can be identified for this scenario. An algorithm will identify location L1 and location L2 as being the same location, yet it would do the same thing in case we ignore AP1 and it would require less time to do so. A bad algorithm might not even consider location L1 and L2 as being the same. This can happen is the algorithm attribute a high weight to the difference between the access points that have been attributed to different locations.

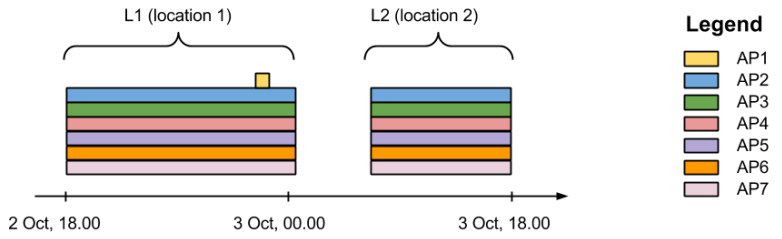

Figure 4.1: Example of an isolated observable access point

The above scenario considers a very small number of access points and a very short period of time. The time gains in eliminating the access point that does not provide so much information in this case would be very small. However, If we are, for example, looking at a month of collected data for a user, we will

have entries for thousands of different access points that were observable at any moment during this time. Out of these entries there can be hundreds of access points which are never visible to the device during close scans and as such their importance when determining the fingerprints for the different locations is very limited, yet they do have a huge impact on the execution time needed to actually extract the locations.

In order to solve this issue, we eliminate from the access points those ones who are not respecting the following condition:

- There is no time window of at least 5 minutes throughout the time duration of the analyzed data in which the access point has appeared for more than 5 times

We have chosen to use time windows of 5 minutes as we make the assumption that any user will choose to spend minimum 5 minutes at each stop location. In case the user spend less time, we can consider that they are just transitioning until the next stop location.

Chapter 5

Extracting locations from Wifi data

Human mobility has been attracting a high degree of attention from numerous study fields among which we find urban and traffic planning, traffic prediction, the spreading of diseases and many others [AGB13] [DB08].

The studies that have been conducted on this subject have been using various ways to identify the travel behaviour of people. Some of them have focused on studying the information gathered from observing the way in which money is dispersed through time [DB06], or they have been focusing in studying the behaviour of mobile phone users by analyzing the way they move based on the telephone towers their phones are connecting to when they are engaging in voice communication [MCG08]. There are studies that try to understand human mobility through the glass of social networks [YYZS10], as it can be observed that individuals prefer to meet with other people that are part of their community more often [MM07]. GPS data has also been considered for various studies [CLL14], [ZG10]. The list of elements that have been taken into consideration for trying to understand and predict the way in which we are conducting our daily travels is far from being short.

The present chapter focuses on the steps of our work that have been taken in order to identify and extract locations from Wifi data. We have analyzed the data in order to explore ways in which locations can be determined based on

characteristics that can individualize them. We have also experimented with various algorithms that can be used in order to extract locations. All these steps are presented in the sections of the current chapter.

5.1 Wifi based positioning

Even from the beginning of the 21st century, research has been actively conducted for trying to use the Wifi system in order to determine real positioning and different databases for positioning systems have been created. These databases usually included the positions of the Wifi access points (APs) or RF (radio-frequency) identified fingerprints [CSC+06] [CCLK05] [YA05] [BP00]. Modern databases for Wifi positioning are created with information about the signal strength for the Wifi APs and can even have information about where they were discovered.

Koo et. al. [KC11] have explored an algorithm that can help estimate the relative positions of APs corresponding to the real geographic configuration with the help of multidimensional scaling techniques. Considering the fact that APs are not able to tell real distances between themselves and other APs, the study aims to estimate the dissimilarities between different APs using scans. They have also conducted an experiment in an office building in order to test the proposed algorithm and the results showed an estimation error of approximately 7 m.

Another study conducted in this similar direction is the one by Mok et. al. [MR07]. The authors explore the possibility of determining the location of a device which can scan Wifi APs based on the signal strength that the access points are displaying at the moment of the scan. They estimate the positioning by performing a trilateration based on the information the device gets from multiple access points. The accuracy for their algorithm for the conditions that were presented in their experiment was of about $1 - 3$ m.

Athanasiou et. al. [AGGP09] give a very clear and concrete description for two classes of wireless positioning systems. Their work focuses on experimenting with parameters for these algorithms in order to find the optimal solution in terms of accuracy under realistic settings. They also adapt a global map matching algorithm in order to extract travel time maps from wireless data and they propose a demonstration for showing that for high sampling frequencies, the locations identified are comparable to the ones derived from GPS data.

The two classes of algorithms that are explored by the authors are: centroid

and fingerprinting. *Centroid* is presented as the fastest method for positioning, however it depends on having the real location of the APs. This information is in general unavailable and as such a proposed solution is to estimate the locations of the access points by calculating an arithmetic mean of all the coordinates at which it was visible. The *fingerprinting* method is based on the assumption that the APs are stable over time (they do not change positions). This leads to the fact that at any time, a measurement at a particular location will return the same list of APs with the same signal strengths. As such, this list can be considered as the unique fingerprint or signature of the location.

Zhang et.al. [ZF12] propose an algorithm based on fingerprinting for estimating locations. The algorithm takes into consideration the fact that the signal strength from various APs does not necessarily stay constant throughout the time. They propose a way in which a similarity between fingerprints can be calculated in order to determine if two fingerprints are in fact representing the same location.

These are just a selection of works that have been conducted on finding a solution for Wifi based positioning systems. With the growth and improvement of Wifi systems, in time all barriers can be overcome and we could have a positioning system that is as accurate yet considerably cheaper than GPS positioning systems.

5.2 Determining the fingerprint of a location

In order to have a better understanding of the way in which the mobile phone users have been moving throughout the experiment, we need to have an image of the way a given period of time would look based on their Wifi records from SensibleDTU. As it has been presented in Section 4.2, the Wifi data we are using for the present project consists in the following fields: user id, timestamp, SSID, BSSID, RSSI and the context. However, considering the amount of data involved, just by looking through the log files it is almost impossible for us to understand at what moment the user might have reached a location and when they left from it. In order to be able to do this, we have created various visualizations considering different options, different time frames and for multiple users in order to begin to understand what the data can tell us, what we can use, what would we need and what can we discard when moving further to defining what can help identify a location.

5.2.1 Signal strength over time

The first thing that we have tried to visualize was the APs that were scanned by users' mobile phones throughout different periods of time. We have plotted the APs and their registered signal strength for a variety of users in order to see if we notice any patterns in their movements.

In Fig. 5.1 we can see how a day from the life of a random user (referred to as userX) looks like. The day for which we have plotted the data started on a Tuesday at $12:15$ pm and ends the next day right before the same hour. The hourly intervals can be seen on the X axis, while the signal strength values can be seen on the Y axis. The legend contains the top 10 most popular [1] APs that have been scanned throughout the given time.

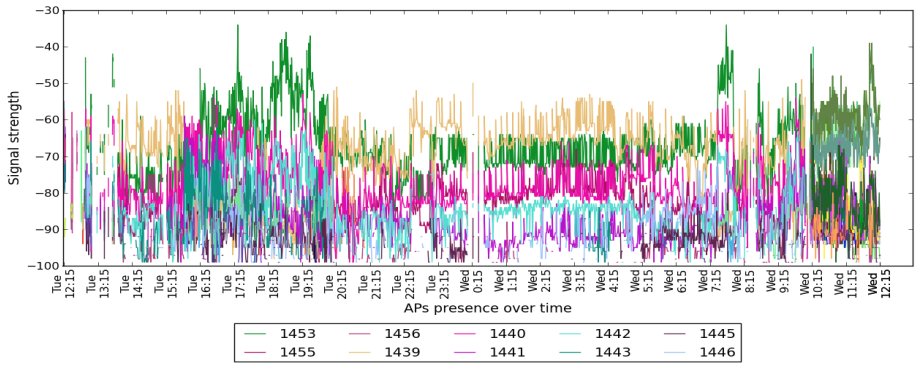

Figure 5.1: Example of the APs registered for an user throughout one day (using connecting lines markers)

The steps for creating this type of visualization are as follows:

- Retrieve data for the time duration for which the visualization is made
- Keep track of all the timestamps at which each AP has been seen and the AP's signal strength at those moments
- In case an AP is scanned no more than 2 minutes after it was previously scanned, then a line can unite the two moments in order to mark their proximity. If the apparitions are more than 2 minutes apart, then there is a high possibility that there has been a location change or that the AP is experiencing technical problems and as such has stopped being active.

[1] An AP is more popular than another in case it appears more times during the period of time for which the Wifi scans are analyzed

Although we have tried to visualize this type of information in various ways (using different types of markers), we found that this way is the easiest to interpret by people [2]. If we leave out the lines, for example, as it can be seen in Fig. 5.2, it is quite hard to interpret where location might start or stop.

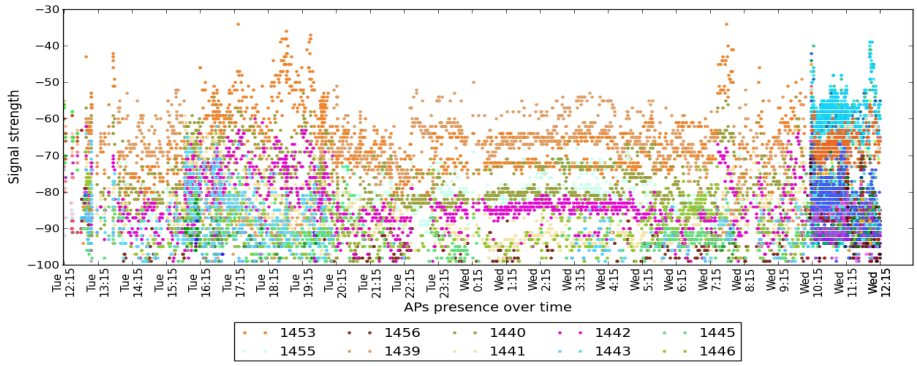

Figure 5.2: Example of the APs registered for userX throughout one day (using point markers)

Other ways in which we have been experimenting with visualization for this can be found in Appendix A.1.

By looking at Fig. 5.1 we are at some level able to distinguish moments of time at which the user seems to be arriving at a location [3], however it is hard to notice any patterns because we are only observing a single day in the life of userX.

Let us look at the data gathered through 7 days from another user's (referred to as userY) life. The visualization for this data can be seen in Fig. 5.3. The image gives out some very interesting information. We can, for example, notice the repeating patterns which are dominated by the orange, light green and blue colors. These patterns appear during the evening and the night and we can assume that the user is spending this time at the location which we can label "home".

We can notice some periods of time that are free. These free gaps like, for example, from Monday morning until Monday evening are gaps in which no

[2] We have required a group of 12 different people to rate different visualizations and one used in Fig. 5.1 has been preferred by a majority of 8

[3] For example, we can say that what we notice from Wednesday at 10:15 until the same day at 12:15 is different than anything we can see before that time so we can assume that it is a new location.

signal was scanned and can mean that either the mobile phone was closed or that the user decided to switch off the Wifi.

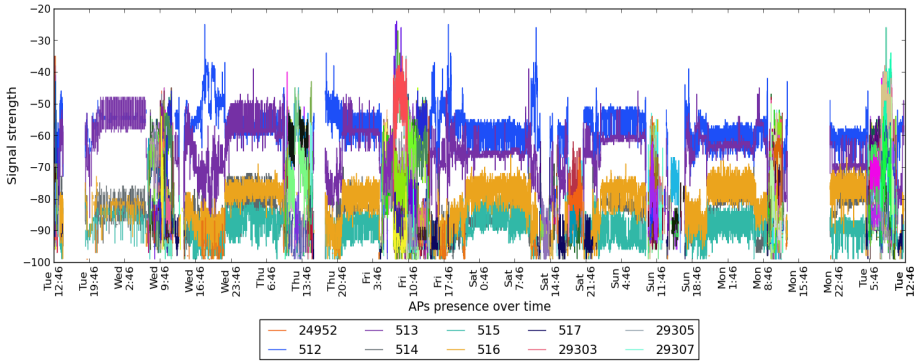

Figure 5.3: Example of the APs registered for userY throughout 7 days

We can also notice fragments in which the density of signals is quite high, for example on Wednesday morning. This means that the user was located in a place which has a large number of APs nearby and since we can notice a regularity in this pattern we can assume that this place can be the University. This might seem unlikely based on the fact that the patterns sometimes is identified during the night, however this particular week is set in October when there are deadlines for school projects that need to be handed in.

As we can see, these visualization can offer us a good first glance at what the locations might be like, yet they also make us consider other things that we can learn about the data. For example:

- How many samples from each AP are received during a given time frame
- What is the average signal for various time frames for a given AP
- What are the running averages for signals from various AP

5.2.2 Sample density

When trying to identify locations based on the Wifi data, it is important to only take into consideration the APs that actively contribute to the fingerprint of the mentioned locations. Before cleaning our data (as it has been described in Section 4.3.2), isolated observable APs can appear and unnecessarily burden the algorithm used for extracting the locations. The best way to identify such APs

is by analyzing the sample density [4] of the samples that are identified during scanning.

In order to determine the sample density for each AP, we need to define a time bin over which the sample density needs to be calculated. We have calculated the density considering a time bin of 5 minutes as we can assume that this amount of time can be considered the minimum duration for which a user needs to be situated in approximately the same place in order for us to not consider that the location is a transition instead of a stop location.

In Fig. 5.4 we have the different APs and their RSSI values at the different moments when the mobile phone has identified them in the scans for userX during the second day of observations. In Fig. 5.5 we can observe the sample density for one of the APs that are predominant during the visualized time frame. As we can see, the number of times the AP is present in the scans throughout the day is quite high and it is registered during numerous different periods during the day. We can easily assume that this AP is one of the key APs that define one of the locations the user has been associated with.

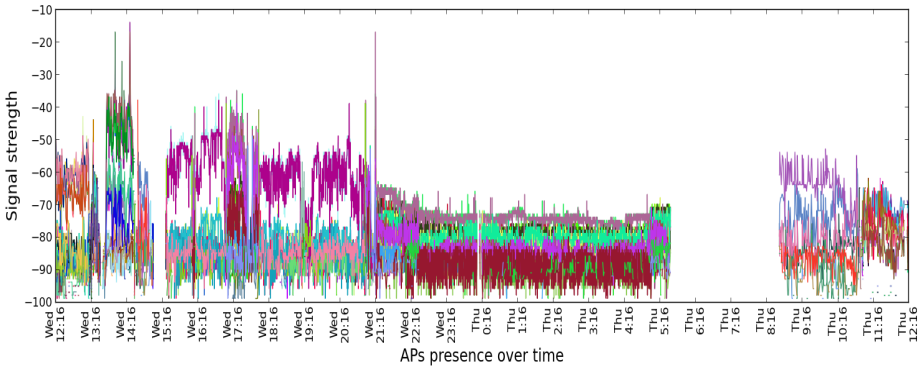

Figure 5.4: Example of the APs registered for userX throughout day 2

On the opposite end as number of times it has appeared during the scans, we have the AP in Fig. 5.6. As it can be seen, this AP only appears 5 times over a one single 5 minute time bin. We can easily presume that the presence or absence of this particular AP will not offer us relevant information over the location at which the user was situated when it appeared in the scans. This statement is also sustained by the fact that the user location seems to be consisted from Wednesday $12:16$ up until around $13:16$ according to what we can observe in Fig. 5.4, even though the AP does not appear throughout most of this time.

[4] We define the sample density for an access point as the number of times it appears in scans over a predefined time bin.

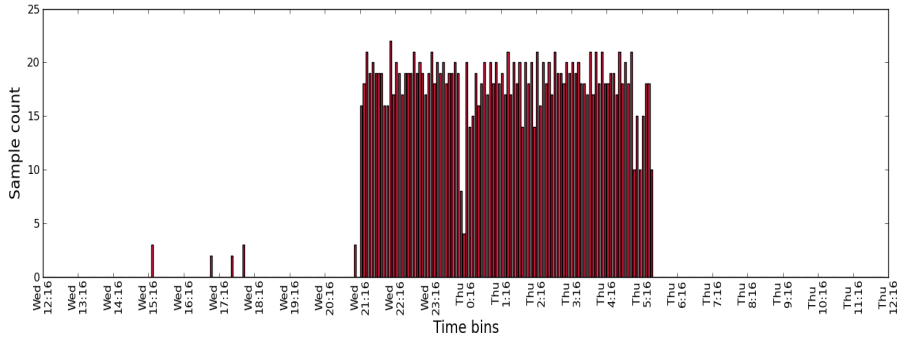

Figure 5.5: Example of an AP which appears often

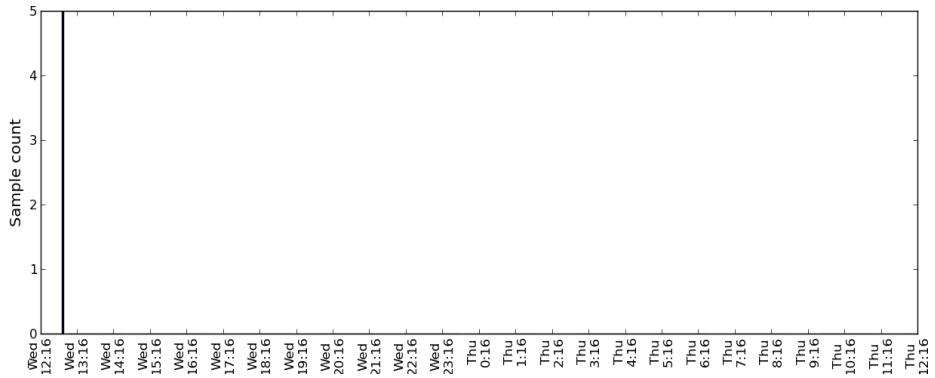

Figure 5.6: Example of an AP that appears just a few times

Other examples of visualizations for APs based on their sample density can be found in Appendix A.2

5.2.3 Exploring the implications of the signal strength

Something that is often taken into consideration during studies regarding the determination of locations based on Wifi data is the value which indicates the signal strength received from the various APs. The level of the signal strength indicator can, in general, give us a good approximation of how close we are to a particular AP. However, Wifi networks are susceptible to interferences [MCWA10], meaning that there are numerous factors which can cause signals to spike even in case the device that scans the region for AP signals does not move.

This can represent a factor of risk when including the signal strength value in the location extraction from Wifi data as the same location could, at different times, be associated to an AP which has a signal strength that oscillates based on other external factors.

In order to see if we can smooth down possible fluctuations we have employed two mathematical tools. We have calculated the average signal strength, as well as the running average for time windows of different length.

5.2.4 Average signal strength

In order to calculate the average signal strength of a given AP for a given time bin, we needed to identify all the moments of time inside the given time bin in which the AP has been spotted during the scans. The average signal of the AP is calculated as the sum of all the strength values that have been recorded for the AP inside the time bin and the sum is then divided to the number of recorded apparitions of the AP. For example, if we were to have an AP which appears 6 times inside a 5 minutes time bin with the following RSSI values [-60, -70, -60, -80, -90, -60], then the average signal strength for this particular time bin for our AP would be $avg = [(-60) + (-70) + (-60) + (-80) + (-90) + (-60)]/6 = -70$ dBm.

We have calculated the average signal for various users and various days. We have also calculated it for different time bin lengths. For example, for the same data that we can see in Fig. 5.4 and for the same AP that has the sample density represented in Fig. 5.5, if we visualize the non-null averages calculated for time bins of 5 minutes, we would have the representation in Fig. 5.7. The X axis records the time while on the Y axis records the values of the averages

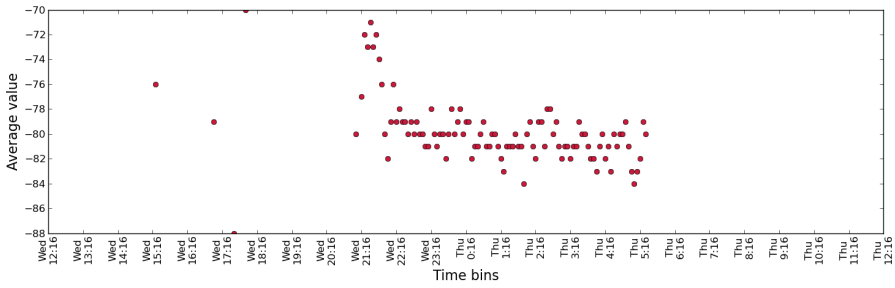

Figure 5.7: Example of average signal strength visualization for userX

The averages are represented by big dots symbols which appear at the beginning

of the time bin for which the average is calculated. For example, if we have calculated an average for the interval $12:05-12:10$, the average is plotted on the visualization at $12:05$.

Additional examples of averages for different APs scanned during a given day can be found in Appendix A.1.

5.2.5 Running average signal strength

The average signal brings a small improvement as far as eliminating the signal spikes go, however, an even better way in order to smooth out any signal fluctuations is to calculate the running average[5] [Hyn09].

We have calculated the running average for different users and time frames, and we have taken into consideration different time bins when calculating it. The algorithm for calculating it is as follows:

- For the selected user and the selected time frame, we have extracted for each AP the timestamps at which it has been identified by the user's phone
- We have divided for each AP the previously mentioned timestamps into bins of 2, 5 or 10 minutes recording also the signal strength identified at each timestamp[6]
- The above identified time bins are overlapping. For example, if a sequence of signals $[-60, -80, -70, -70]$ that have each been identified at 1 minute apart is to be divided into bins of 2 minutes, the resulting 2 minute bins would be: $[-60, -80]$, $[-80, -70]$, $[-70, -70]$
- The running average is calculated as the sum of the values that can be found in a time bin which is then divided to the number of values. For example, for the above time bins, the running averages would be -70, -75 and -70

In Fig. 5.8 we can see the APs associated with another user (referred to as userT) and their signal strengths over a day. Fig. 5.9 shows the signal strength for just one of the identified APs. The average signal as is presented in Section 5.2.4 for the same AP can be seen in Fig. 5.10. Fig. 5.11, Fig. 5.12 and Fig. 5.13 present the running averages calculated for the same AP for time bins of 2, 5 and 10

[5] Also referred to as the moving average

[6] By doing this we have the signal strength for the given AP at any moments it has appeared inside the time bin

minutes [7]. The X axis of these figures tracks the succession of time moments while the Y axis keeps track of the value of the running average calculated over this time.

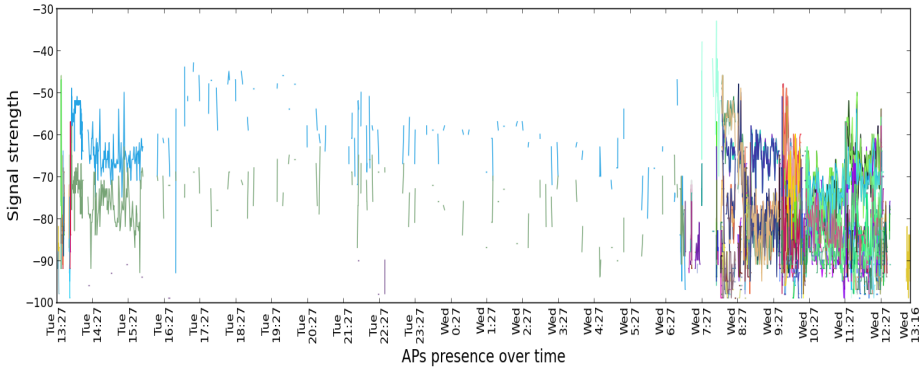

Figure 5.8: Example of APs presence over time for userT

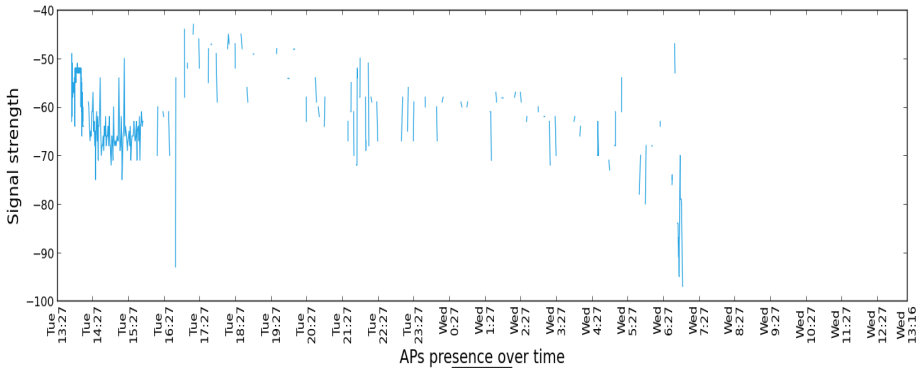

Figure 5.9: AP 85 for userT during 1 day

The way in which the fluctuations are smoothed down can be easily seen in the figures that present the running averages calculated for various time bins. The fluctuations are smoother as the time bin is increased.

Visualizations for running averages calculated for other APs identified during the same day for userT can be found in Appendix A.2.2.

[7] In this representation, only the non-null values for running averages are displayed

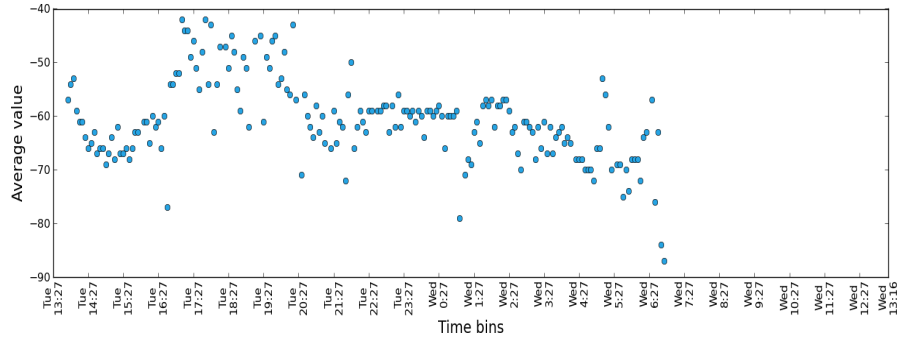

Figure 5.10: Average strength for AP 85 for userT during 1 day

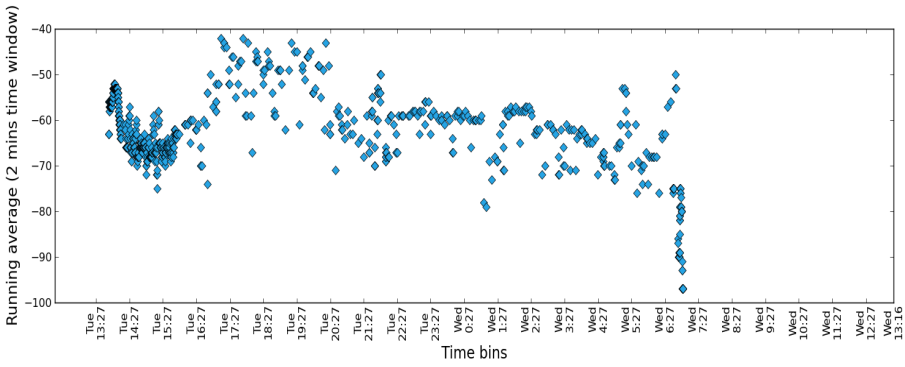

Figure 5.11: Running average for AP 85 for userT during 1 day (2 minute time bins)

5.2.6 Signal presence

Even though averaging the signal strength through time improves at a certain level the fluctuations in the signal strength, in a real environment spikes will always be present and this will bring extra difficulties in estimating locations based on fingerprints that contain the value of the signal strength for the involved APs.

Another way of looking at locations is by calculating their fingerprint based only on the identity of the APs that have been identified while the user was found at that particular location. Basically, instead of defining a location based on both the identity of the APs present and their signal strength, we would only associate locations to visible APs.

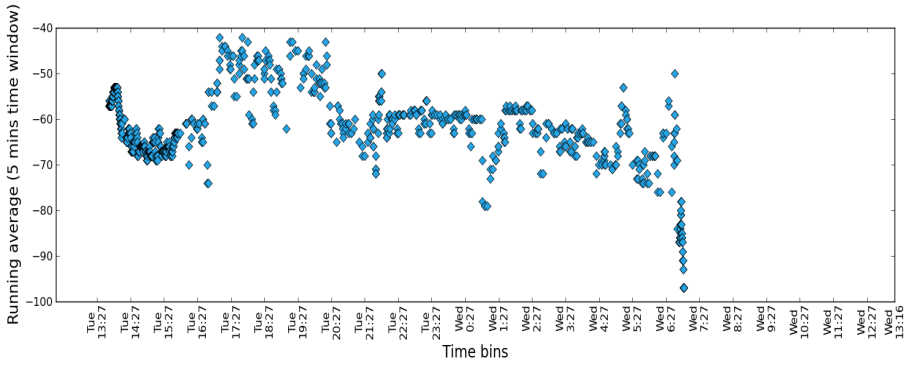

Figure 5.12: Running average for AP 85 for userT during 1 day (5 minute time bins)

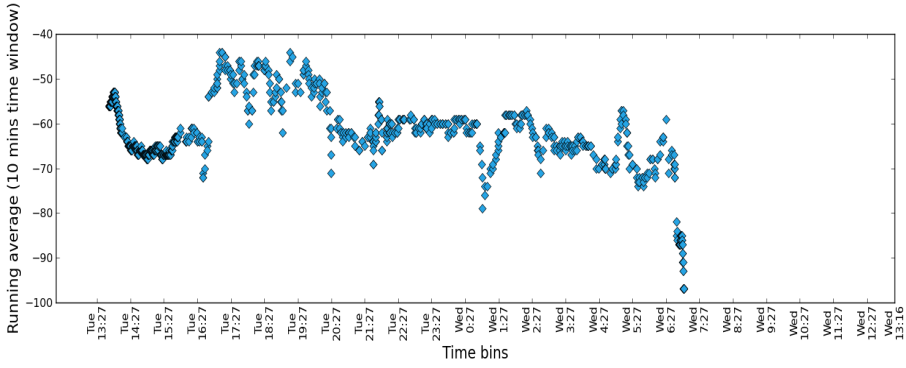

Figure 5.13: Running average for AP 85 for userT during 1 day (10 minute time bins)

The idea is simple and elegant and has been used in previous studies with success [LJ09]. The concept behind it is that, in general[8], at a given location the scans will always show the presence of the same APs. If, after a time, the scans change and other APs appear, it is reasonable to assume that the user has changed locations.

Since the information offered by the signal strength does not seem to be of the ultimate importance, we can, in this case, try to identify the locations only based on the presence of the APs. We consider that an AP is present at a specific

[8]New APs can be set up or old ones can be changed with new ones in time, which would mean a change in how the scans would look for the same location. However, this is an issue that is outside the scope of the present paper and work.

moment of time if the Wifi scans at that moment register a signal strength from that AP. However, as it has been mentioned previously, due to interferences, the signal from the AP might be lost for short periods of time even when the user does not change their location. Considering this and the assumption that, in general, people tend to spend at least a few minutes in a stop location (otherwise meaning that they might be just transiting it), we have made the decision to adapt for our case the definition for the presence of an AP.

We divide our data into time bins of 5 minutes [9]. We redefine the presence of an AP as follows: an AP is considered to be present for the duration of a 5 minute time bin if it appeared in the scans at any point inside this time interval.

We can use visualization in order to see how this transforms the way in which we can understand the data. In Fig. 5.14 we have the different APs that have been scanned throughout the duration of 2 days for userX. In Fig. 5.15 we can see the top 50 predominant APs and their presence over 5 minutes time bin during the same 2 days [10]. The X axis keeps track of the time bins throughout the 2 days, while the Y axis represents the annonymized identifiers for the APs.

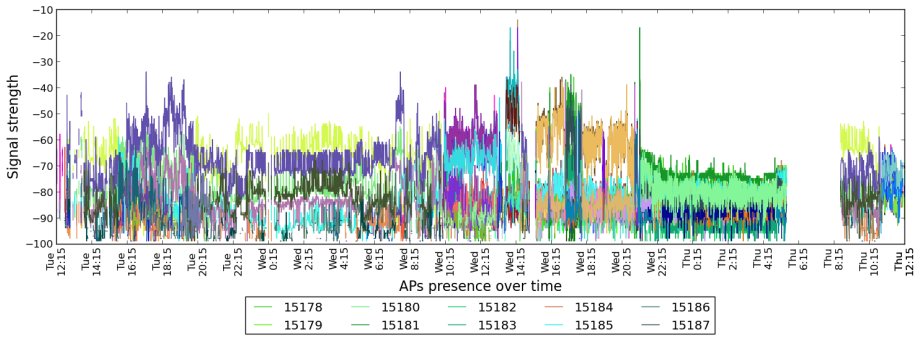

Figure 5.14: Scanned APs for userX throughout a duration of 2 days

By closely observing the two visualization, it is quite easy to see that indeed they are representations of the same period of time. Even if not all APs are displayed in the visualization for the presence of the APs over time, we can notice that, for example, the user has spent the time from Wednesday 21 : 15 until almost Thursday 6 : 15 in one location. This also coincides with what we can observe in the visualization for all the APs (with signal strength) scanned

[9] We consider 5 minutes as the minimum amount of time that needs to be spent in a location for it be considered a stop location. This number can be easily adjusted in case further research shows that it is not the optimal value to be taken into consideration.

[10] We restrict our visualization to 50 APs as it would be hard to understand an image in which we would be displaying all the hundreds of APs which were encountered throughout the 2 days.

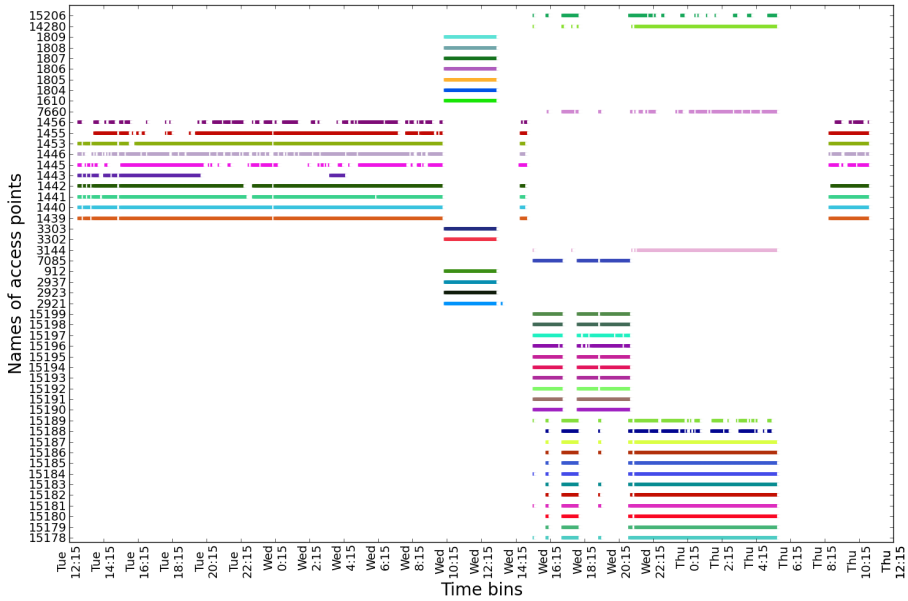

Figure 5.15: The most common 50 APs for userX during the given 2 days (presence visualization calculated for 5 minutes time bins)

throughout this time.

In Appendix A.2.3 can be found a visualization for the presence of APs for a period of 2 days for another user. The presence for APs is determined for 5 minutes time bins over the 2 days.

5.3 Extracting locations

By visualizing the Wifi data in the way presented in Section 5.2.6, we can begin to see how locations seem to succeed each other throughout the days for a particular user. However, it is important to be able to implement a solution that will extract these locations from a large amount of data so that we would not be needing to examine the data manually. We have used different methods in order to get the best possible approximation for identifying the locations. The methods we have tried are: using *networks*, using *k-means clustering* and using *Hidden Markov Models*.

Before describing each of these methods, we have to clarify what we consider a fingerprint of a location at a given time. A fingerprint of a location is calculated based on the AP presence inside 5 minutes time bins (Section 5.2.6) as follows:

- We extract the data we want to analyze from the user (either for 1, 2 or more days).
- We identify the APs from the data
- We divide the data into 5 minutes time bins
- For each time bin we identify the APs which have been spotted during the 5 minutes and we attribute them the value of 1 (meaning that they are visible to the mobile device during that time bin); the remaining APs will have attributed the value 0 (not visible) for the given time bin.
- Each fingerprint describes a time bin and shows what APs are visible during it and which are not

The fingerprint contains the names for all the APs which are associated to the user throughout the time frame we are analyzing (for example 1 day, 5 days etc.) and each AP has associated to it a value which represents its presence throughout the 5 minutes.

5.3.1 Network theory

Networks have a high degree of importance when trying to understand human or animal behaviour. They have been used in combination with social platforms in order to extract a new definition of friendship [CML11], they have been used for monitoring animal behaviour [GCP+06], or to understand the economical situations caused by the way in which people interact [CJK05], or just to understand underlying communities of people. During our research, we have considered the use of network theory in order to extract locations from Wifi data.

In theory, we can expect that the APs that are identified at a particular location will not appear in the scans the user's phone will have from another location. This assumption is sound as the APs will rarely be moved and as such they should always be associated to the same place. In this case, we should expect that the succession of locations can be similar to what we can see in Fig. 5.16., where AP1-AP6 are associated to location number 1, while the remaining APs are associated with location 2 and the APs never overlap.

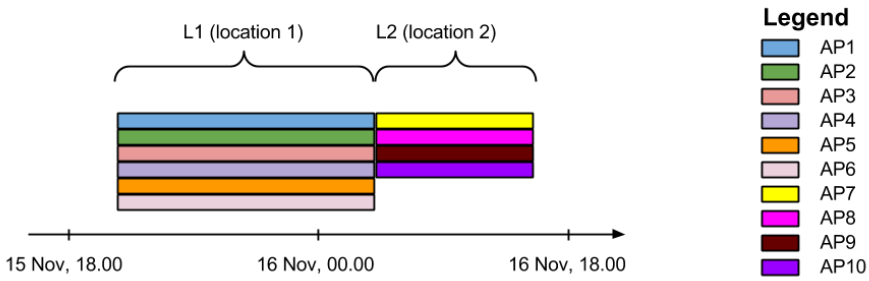

Figure 5.16: Example of how, in theory, locations should be displayed through the presence of APs

The idea behind constructing the network that can be used to extract the locations is simple. A graph can be created for each user from their data and the locations can be identified as follows:

- We consider each found AP from the user data as a node in the created graph
- We construct a presence matrix for the identified APs. Each line in the presence matrix is associated to an AP and contains the signal presence (Section 5.2.6) calculated for the AP considering 5 minute time bins throughout the time we are evaluating
- For each time bin, we identify the APs that are present throughout it and we connect each two of them with an undirected edge
- Since signal from various APs can be lost due to interferences, for each two APs for which we have created a connecting edge, we keep and update a variable which represents the number of times the APs have been identified in the same time bin
- After the network is completely created, we normalize the counts of how many times each two APs have been seen in the same time bin by dividing the counted value to the maximum number of apparitions of either of the two access points
- After the normalization we remove the weak links [11]

[11] In this case, a link is considered weak if after normalization its associated value is below a given threshold

- We consider the resulting connected components to be the extracted locations

An example on how to construct such a graph if given 4 APs and their presence matrix can be seen in Fig. 5.17

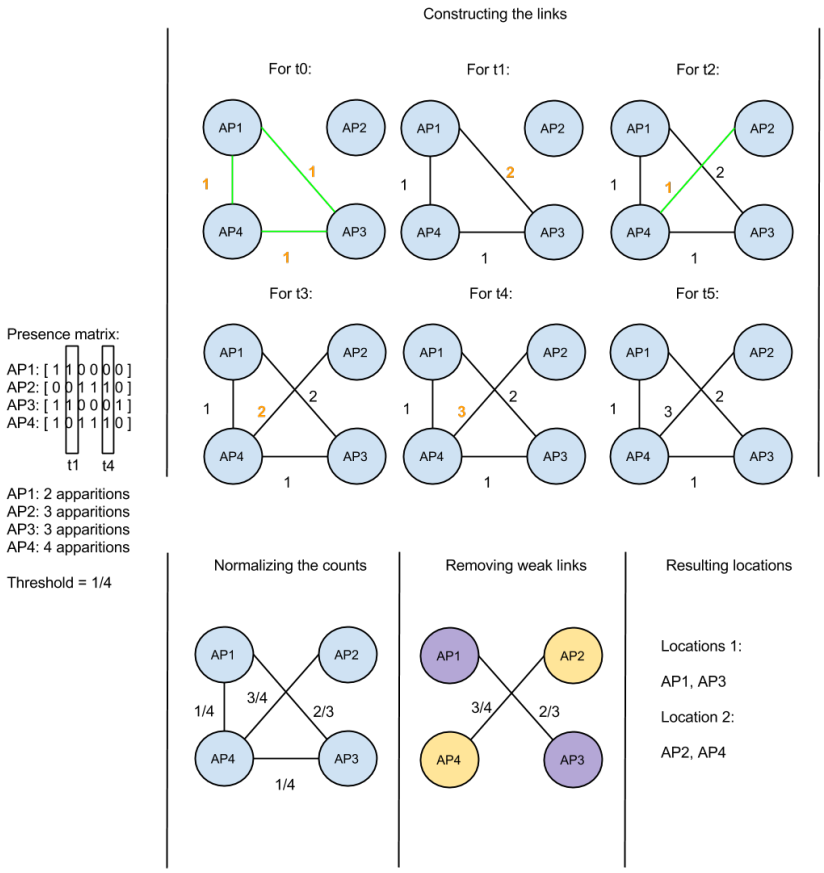

Figure 5.17: Example of constructing a network

We have applied the previously described algorithm for a selection of users, but the results have not been satisfactory.

For example, we can take data for one day for userX. The visualization for the identified APs and their presence throughout this time can be observed in

Fig. 5.18. By looking at this image, we can observe that the user has been in f2 main locations during this day.

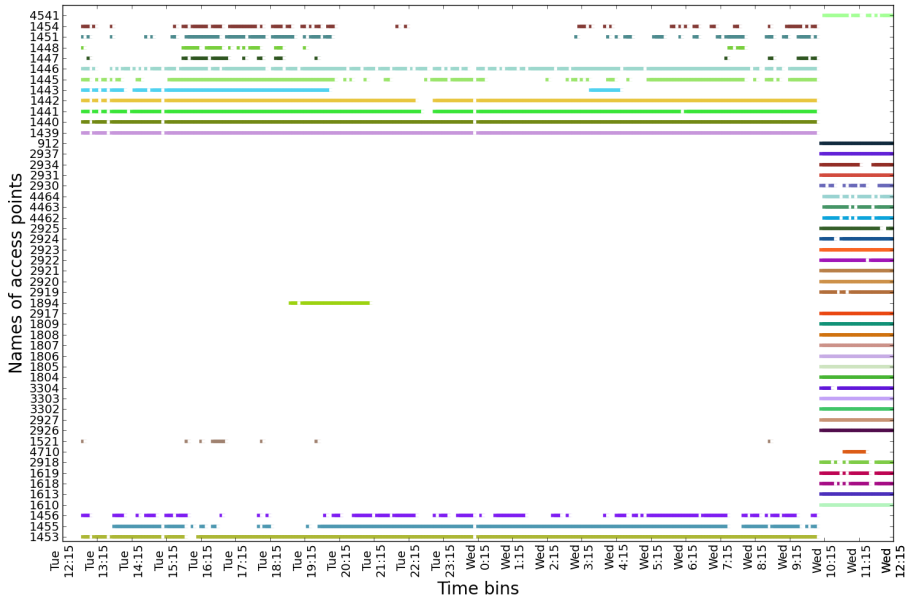

Figure 5.18: The most common 50 APs for userX during one day scan records

When running the algorithm that extracts the locations based on the constructed network, we obtain 21 connected components which have the potential of being locations. The image for the connected components can be seen in Fig. 5.19

We have tried to adjust the threshold for eliminating weak links yet the results are in most cases unsatisfactory. Upon closer analysis we have observed that this can be caused by the fact that, sometimes, there are a high number of APs that even though they are in reality tied to a given location, their signal fluctuates often and as such, the algorithm identifies them as part of a different location. As such we have locations consisting in only a very small number of APs that in reality could have been integrated in other locations. This observation is sustained by the size distribution of the generated networks (example for such a size distribution determined for the network in Fig. 5.19 can be seen in Fig. 5.20). Another thing we have observed is that there are adjacent locations which can have interfering APs signals. This means that our original supposition that, at all times, the APs that are visible from a location will stop being visible in any other location does not always hold.

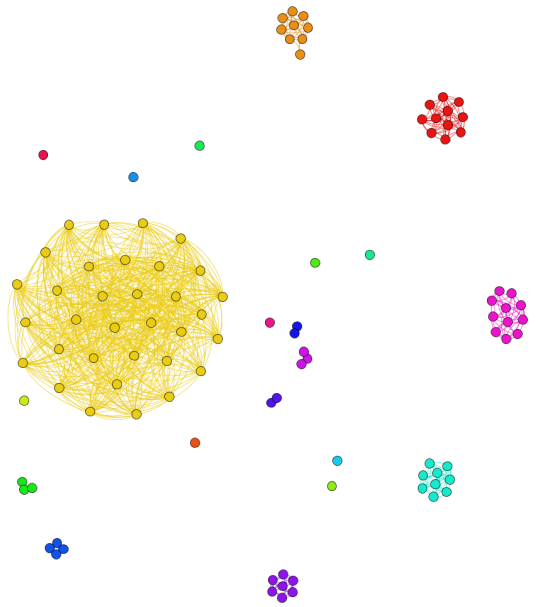

Figure 5.19: Locations identified with networks for userX during one day

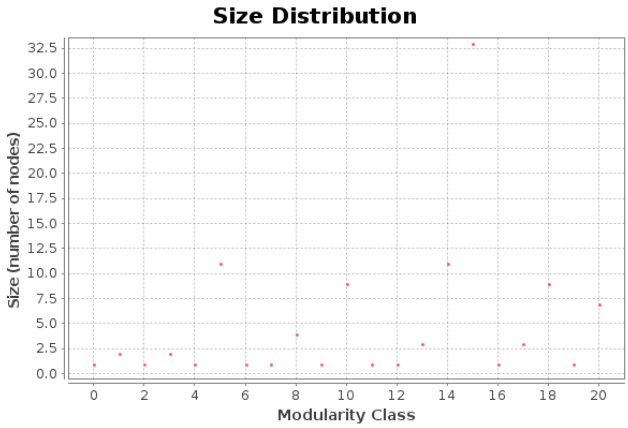

Figure 5.20: Size distribution of locations based on the number of APs associated to them

For generating the networks and the size distributions, we have been using Gephi [Gep]. For obtaining the visualization in Fig. 5.19 we have used the Force Atlas 2 layer which has been configured in order to avoid overlapping of

components.

5.3.2 Cross validation

For both the k-means and the Hidden Markov Models approaches on extracting locations out of the user data we have been faced with a problem. The problem is that both these algorithms need to know how many locations they are trying to identify. However, we cannot know for sure, from the beginning, how many locations a user has been visited during a given time. In order for us to have a good estimation for the number of locations we could be expecting to find inside a time frame, we have used the cross validation technique, more specifically we have used the 10-fold cross validation method.

Cross validation [Koh95] is a technique for model validation that tries to asses how the results given by a statistical analysis of some given data can be generalized to an independent data set. The main use of cross validation is in problems that deal with prediction. Prediction problems usually deal with a set of training data and a set of testing data that the model needs to be able to react to as expected.

The k-fold cross validation divides the data we have at our disposal in k equal sized subsamples in a random way [12]. k-1 of the resulting subsamples are used as training data, while the remaining subsample is used as testing data. The samples are then rotated so as each of them becomes, in turn, testing data while the others form the training data. The k results are then combined in order to retrieve an unique estimation for the original data. An evaluation of the accuracy of the prediction model can be done based on how close the result is to the original data. The k value can be any number as long as the data can be divided into k subsamples. A value that is often used for k is 10 [MDA05].

An example of how 2-fold cross validation works for 4 location fingerprints can be seen in Fig. 5.21.

In Step 1 we have the four fingerprints. Let us consider that the algorithm that we are using to extract the locations based on these fingerprints has identified locations 1 and 2 as they can be seen in Step 1. In order to see if our algorithm behaved as expected, we can cross validate the result using in this case a 2-fold cross validation. We are randomly selecting the 2 subsamples as is seen in Step 2. The blue color is associated with the training data, while the orange color

[12] Random in this case means that each of the subsamples contains elements from the original sample that are most likely not in their original order.

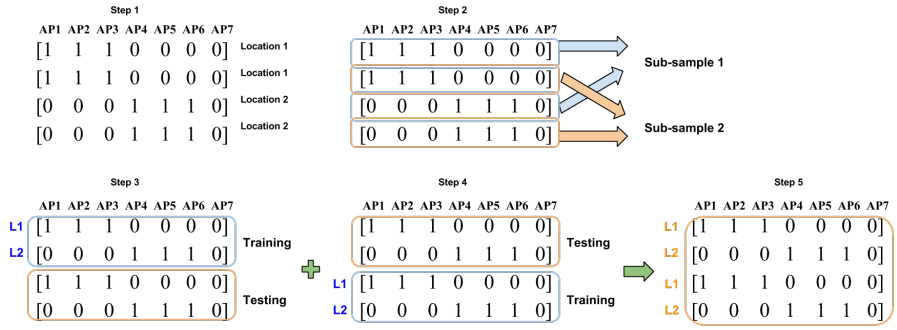

Figure 5.21: Example for 2-fold cross validation

is associated with the testing data. In Step 3 the first subsample is treated as training data while in Step 4 the second one represents the training data. After the test data is classified based on the training data, we can combined the results into one single sample which is presented in Step 5. In this example, the cross validation has returned a result which matches the original estimation made by our selected location extraction algorithm, which means that the algorithm we have used has worked properly.

5.3.3 K-means clustering

The k-means algorithm is a popular method for analyzing clusters in data mining. The first time the "k-means" term was used was by MacQueen [Mac67], yet the standard algorithm for the k-means problem was proposed by Lloyd [Llo06]. The idea behind this algorithm is to start the clustering process by having k original groups of only one point each. After this initial setup, each new point can be added to the cluster that has the mean nearest to it. After a new point is added to a group, the mean of the group is updated, and at each stage the k means will represent the means of the k groups.

The Lloyd algorithm for k-means is the solution that is used for creating the k-means tool by scikit-learn [SL]. We have used the tool provided by scikit-learn in order to try to extract the places where a user has been situated at during a time frame based on the fingerprints that we can determine for that time frame. Our goal was to use the k-means algorithm to cluster the fingerprints that are similar enough as to be associated to the same location. However, since we cannot know for sure the number of locations (clusters) we are to expect for a given time frame, we are running the k-means algorithm with different values

for k and we perform 10-fold cross validation in order to see what value has generated the most likely estimation.

The steps in extracting the locations with k-means and 10-fold cross validation are as follow:

- We select the time frame (number of days) for which we want to extract the locations
- We retrieve the data and extract the fingerprints
- Since previous research shows that in general people spend most of their time in a small amount of locations (5 to 50) [MCG08], we choose the maximum number of locations we are expecting to find as the minimum value between 50 and the result for the number of days multiplied by 10[13]
- For each possible number of locations from between 2 and our previously selected maximum we run the k-means clustering algorithm on the identified fingerprints
- The algorithm returns the estimations for the locations. The locations are the k clusters formed based on the differences appearing in the presence values associated to the APs in the given fingerprints. These differences determine the distance between the fingerprints
- The estimations are cross validated in order to see which number of locations has generated the optimal approximation for the given fingerprints
- In case more locations have generated equally good results we selected as number of expected locations the highest of them
- This algorithm is ran 10 times leading to 10 estimations for the number of locations. Out of these estimations the one which appears the most times out of the 10 results is considered correct.

We run the algorithm for 10 times in order to ensure that the final result is as less influenced by the random fact involved in the determination of the subsamples for the cross validation as possible. At the end of the algorithm we have the estimation for the locations throughout the selected time frame. An example for such an estimation can be seen in Fig. 5.23, while in Fig. 5.22 we have the presence of the APs which have been scanned throughout the same amount of time for the same user.

[13] By observing the visualization for the presence of APs during different days in different users' life, we have observed that in general they seem to spend their time during a day in at most 10 locations

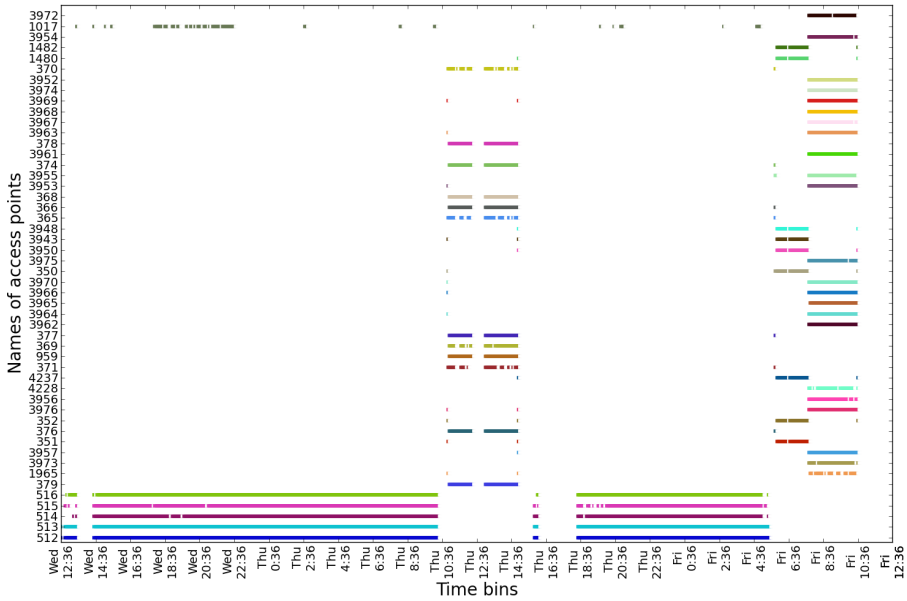

Figure 5.22: The most common 50 APs for userY during the given 2 days (presence visualization calculated for 5 minutes time bins)

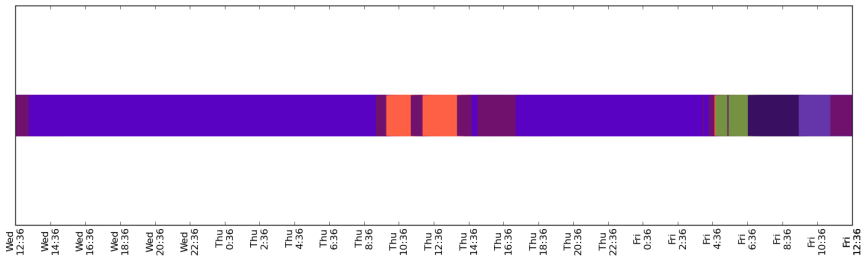

Figure 5.23: Locations estimated with k-means for userY for 2 days

We have evaluated the results of the algorithm in a graphical manner. We have generated visualizations for the presence of the various APs (like in Fig. 5.22) and made a graphical analysis of how similar the patterns are to the results given by the k-means algorithm. The algorithm identifies a big part of the locations, however differences do appear. For example, in Fig. 5.23 we can see that on Friday at $10:36$, according to the k-means algorithm the user has been in a location that changed at some point before $12:36$. This observation is inconsistent to the image we have based on the APs presence in Fig. 5.22.

An additional example of locations estimated using k-means can be seen in Appendix A.2.4.

5.3.4 Hidden Markov Models

"Hidden Markov Models (HMMs) are a formal foundation for making probabilistic models of linear sequence 'labeling' problems. They provide a conceptual toolkit for building complex models just by drawing an intuitive picture. They are at the heart of a diverse range of programs, including genefinding, profile searches, multiple sequence alignment and regulatory site identification. HMMs are the Legos of computational sequence analysis."[Edd04]

HMMs are, in principle, Markov Models (MMs) [Dra67] for which the modeled systems are considered to be processes with hidden states. If in the MMs the states are visible to possible observers, the difference with the HMMs is that the states are not visible, yet the results which can be observed do depend on the hidden states [Rab89].

There area a few elements that characterize the HMMs according to [Rab89]. They are as follows:

- N which represents the number of hidden states in the model. In general, these states can be interconnected in a way so that from some states others can be reached
- M which represents the number of different observation symbols that are generated by the hidden states
- A which represents the state transition probability distribution. A is a matrix for which each element $a_{i,j}$ represents the probability of moving from the state i to the state j in the system represented by the HMM
- $B = b_j(k)$ which represents the observation system probability distribution in state j. This basically means that each element in B shows the probability of seeing a particular element of M in a given state j
- π which represents the initial state distribution, meaning what is the probability of the system to start producing output from any of the states in N

The three problems also mentioned in [Rab89] that the HMMs can be used for solving are as follows:

- Given an observation sequence, how can the probability of the observation sequence be computed efficiently considering the given model?
- Given the observation sequence how can a state sequence be chosen so that it explains in the most appropriate manner the existing observations?
- How can the parameters of the model be adjusted in order to maximize the probability of a given observation sequence?

The second of the three problems above addresses the uncovering of the hidden states of a given model. This can be used in our case because we need to identify the locations an user has been at based on observing transitions between different fingerprints of the locations at given times and without knowing what those locations actually are. In our case, N represents the number of unknown locations a user has been at, M is the set of observable fingerprints which we can calculate based on the presence of various APs in 5 minutes time bins, and A, B and π are the various probability distributions that can be associated with the way in which the user travels from location to location.

The idea of using HMMs in order to track localization is not new. It has been explored in papers like [EKHH13], [IHO13] or [MNRS07] which sustain the potential of using an algorithm based on this method for studying the travel behaviour of people.

Scikit-learn [SL] offers an implementation for HMMs that ensures the training for the models and the inferring of the hidden states and we have been using the tools they provide for working with our data.

As with the k-means method (Section 5.3.3), the problem we have been facing is that we can not approximate from the beginning the number of hidden states (which stand for locations in our case) that we are expecting the model to find based on the input observations. However, by using the k-fold cross validation (Section 5.3.2) we can, once again (similar to the way in which we have solved the problem for k-means), test the estimation being computed based on different numbers of possible locations.

The steps in extracting the locations with the help of the HMM based algorithm that has been combined with a 10-fold cross validation are as follow:

- We select the time frame (number of days) for which we want to extract the locations
- We retrieve the data and extract the fingerprints

- We choose the maximum number of locations we are expecting to find in a similar manner we have done for the k-means algorithm, meaning as the minimum value between 50 and the result for the number of days multiplied by 10
- For each possible number of locations in the range of 2 and our previously selected maximum we run the HMM algorithm on the existing fingerprints
- The estimations are cross validated in order to see which number of locations has generated the optimal approximation for the given fingerprints
- In case more locations have generated equally good results we select as number of expected locations the highest of them
- This algorithm is also ran 10 times leading to 10 estimations out of which the one which appears the most times out of the 10 results is considered the correct one

The reason behind running the algorithm 10 times is the same as the one presented for the k-means algorithm. We want to ensure that the random factor which is involved in the cross validation process has a very little effect on the correctness of the estimation. At the end of the algorithm we have the hidden states (in our case, locations) that can be extracted based on the observations we have based on the presence of the various APs the user is associated to throughout the given time frame. An example of locations that have been found for userT throughout 1 day can be seen in Fig. 5.25. They can be easily mapped to the locations we can observe by looking at Fig. 5.24 where we have the presence of the APs which have been scanned throughout the same amount of time for the same user.

We have done a graphical analysis of the results provided by the HMM algorithm considering the results obtained for different users and we have observed that the results tend to be more accurate that the ones obtained when using the k-means algorithm. Fig. 5.25 and Fig. 5.24 illustrate a very good identification of locations done by using HMM.

An additional example of locations estimated using the HMM based algorithm can be seen in Appendix A.2.5.

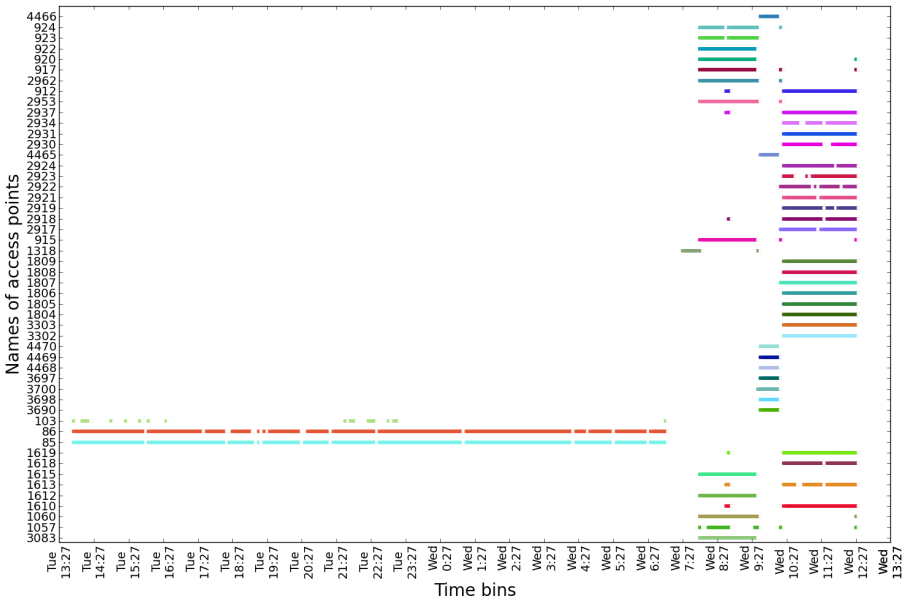

Figure 5.24: The most common 50 APs for userT during the given day (presence visualization calculated for 5 minutes time bins)

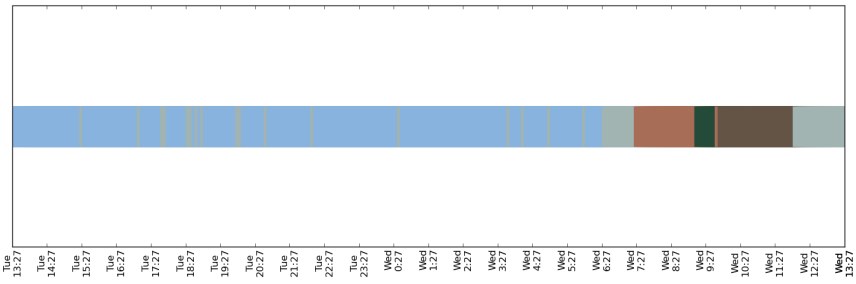

Figure 5.25: Locations estimated with HMM for userT for 1 day

Chapter 6

Location matching

The HMM method as well as the k-means method offer us with the possibility of extracting locations over a given number of days. However, both the algorithms perform better when the given time frame is shorter. This finding has come up while carefully observing the results for different number of days for which the algorithms have been executed.

The reason behind this behaviour seems to be the limitation of the 10-fold cross validation which has been used in order to evaluate the fitness of the results. The method implies that the data from the time frame taken into consideration is divided into 10 equal subsequences which are afterwards used in turns as training data and as testing data. However, when the data size grows, the randomly divided subsequences also grow. When we are dealing with subsequences which have a considerable size, it can happen that some of the subsequences can contain all of the fingerprints which can be attributed to a certain location and as such, that location cannot be estimated based on the other subsequences which have no knowledge of it. This leads to a decay in the efficiency of estimating the number of locations that we can expect the user to have been at throughout the evaluated time. The current chapter explores some possibilities that can help solve this problem.

6.1 Methods for solving the "matching of locations" problem

A solution for the previously mentioned problem can be to scale the k factor of the k-fold validation in order to use a factor larger than 10 when dealing with bigger amount of data, however this leads to a very long processing time which can be avoided by using another solution. The second solution is to extract locations for each day and concatenate the results for all the days afterwards. This however leads to a new situation. We need to find a way in which to identify that a location L_x from day X might be the same as a location L_y from day Y. This problem is referred to in the present paper as the "matching of locations" problem.

We have taken into consideration three possible ways in which we can solve this new problem. In order to evaluate the proposed solutions we have evaluated the results using a graphical approach. We have employed the solutions for a selection of users for whom we have used the HMM algorithm to identify locations through a large number of days. Using each proposed solution we have matched the locations throughout the days and we have observed the accuracy of the matching that was done.

6.1.1 Dictionary of locations based on APs

The algorithm which can be used for matching up locations over time based on a *dictionary based on APs* is as follows:

- For each location we reunite the time bins that has been associated with it. We identify the APs present in either of the time bins and we consider all these APs to be associated to the given location
- Before adding a new entry in the dictionary [1], we can first check the dictionary for previously defined locations that seem to resemble the new one based on the APs that define them
- If we do not find a similar location we can just add a new entry for the new location in the dictionary
- If we find locations which resemble the new one in a proportion bigger than a given threshold (they have a sufficient number of APs that coincide with

[1] Adding an entry in such a dictionary is equivalent with defining a new location based on the APs of which presence the new location is characterized by.

the ones associated to the new location), we can chose the one which resembles the most and consider that the new location and this specific location are the same

- In the above case, the entry for the location in the dictionary is updated to contain the reunion of the APs which define the previous definition of the location and the newly found location that matches it

This method seems to be a simple solution, however, at a close exploration of possible situations which can appear within our data we identified a case in which this solution fails. Let use explore the situation in Fig. 6.1.

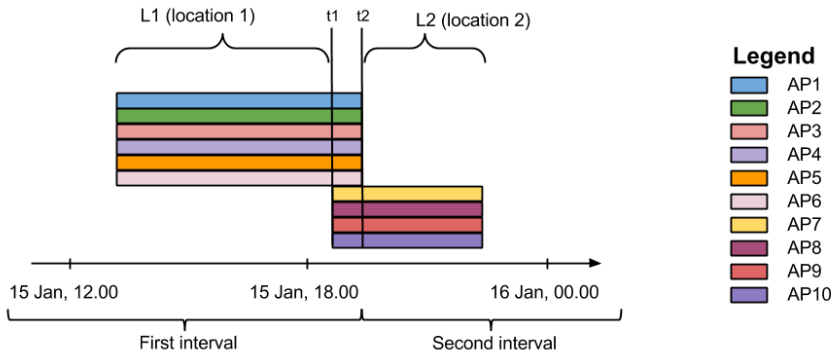

Figure 6.1: Example of APs overlapping

In this case the user would spend most of his or her time for the exemplified duration at location L1 or at location L2. However, these locations can be, for example, two rooms which are close enough so as to have a signal overlap between them (e.g. between t1 and t2). If the user is to stop for enough time, but not very long (e.g. 5 minutes which is the duration of exactly one time bin) in an intermediate location where the signals from the APs in the two rooms overlap, then the extraction algorithm would identify the location as one of either L1 or L2. When using the matching algorithm, the dictionary will contain an entry for L1 which has all the APs from 1 to 10 as all of them have been identified throughout the time in which the user seems to be situated at location L1. Because of this, when the algorithm tries to see if location L2 can be matched with another previous location, since the APs attributed to L2 are AP7-AP10, then all of them can also be found in the entry existing for location L1 and as such the algorithm considers that locations L1 and L2 are the same.

Situations like this one make the use of this particular algorithm to be inefficient.

6.1.2 Dictionary of locations based on fingerprints

The algorithm which can be used for matching up locations over time based on a *dictionary based on fingerprints* is as follows:

- For each location identified with our location extraction algorithm and for each time bin in which the location appears we can see which is the fingerprint [2] for the given time bin
- An entry in the dictionary can contain the fingerprints which are extracted from the time bins that are associated to the location for which the entry is created
- Before adding a new entry in the dictionary, we can first check the dictionary to see if any previously defined location might fit the characteristics of the new location we are trying to add [3]
- If a similar location is not found, we can proceed with adding a new entry in the dictionary for the new location
- If we find locations which resemble the new one in a proportion bigger than a given threshold (which happens when they have sufficient fingerprints in common), we can chose the one which resembles the most and consider that they are the same
- In the above case, the entry for the location in the dictionary is updated to contain the reunion of the fingerprints which are attributed to the previously defined locations which are now matched into one

This solution eliminates the problem that was found in Section 6.1.1 since, in this case the entries for location L1 and L2 would be different as it can be seen in Tab. 6.1 [4].

There is, however, another situation which we were able to identify and which creates difficulties for the good functionality of the present algorithm. The mentioned case can be seen in Fig. 6.2

[2] A fingerprint is calculated for each time bin. It is a list with N elements, where N is the number of APs which are associated with the given user. Element at position i in the list is attributed to AP_i. Each element can be either 0 or 1 marking the absence or presence in the given time bin of the AP which corresponds to the element.

[3] Meaning that a high number of fingerprints are common to both locations

[4] We make the assumption that the algorithm for extracting locations associates the period between t1 and t0 to location L1.

	Location 1 (L1)		Location 2 (L2)
Access points (APs)	Fingerprint 1	Fingerprint 2	Fingerprint 1
AP1	1	1	0
AP2	1	1	0
AP3	1	1	0
AP4	1	1	0
AP5	1	1	0
AP6	1	1	0
AP7	0	1	1
AP8	0	1	1
AP9	0	1	1
AP10	0	1	1

Table 6.1: This table shows the fingerprints for locations L1 and L2 in Fig. 6.1

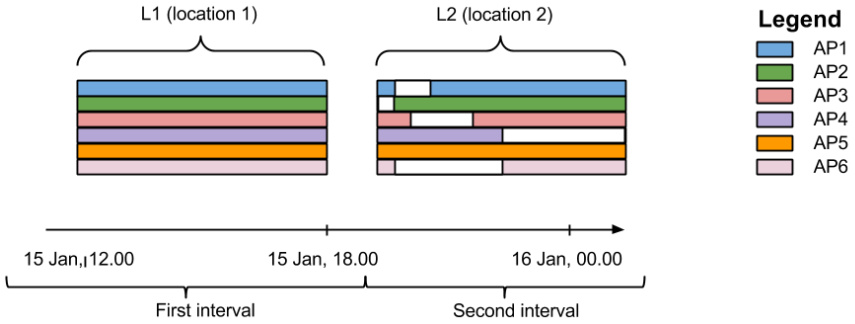

Figure 6.2: Example of locations which can be matched yet do not have any identical fingerprints overlapping

The fingerprints of locations L1 and L2 in this new case can be seen in Tab. 6.2. As we can see there are no common fingerprints identified for the two locations because in the case of location L2 the signal from the APs that identify with it is not constant because of possible interferences. In this case the algorithm will fail to identify the two location as being the same one and, as such, the present algorithm can present difficulties when solve the matching problem.

6.1.3 Dictionary of location signatures

We define the signature of a location as follows:

	Location 1 (L1)	Location 2 (L2)					
APs	**Fingerprint 1 (FP1)**	**FP1**	**FP2**	**FP3**	**FP4**	**FP5**	**FP6**
AP1	1	1	0	0	1	1	1
AP2	1	0	1	1	1	1	1
AP3	1	1	1	0	1	1	0
AP4	1	1	1	1	1	0	1
AP5	1	1	1	1	1	1	1
AP6	1	1	0	0	0	1	0

Table 6.2: This table shows the fingerprints for locations L1 and L2 in Fig. 6.2

- It is an entity which is calculated for a location taking into consideration a given period of time (for example 1 day) for which we have used an algorithm for extracting the locations
- It is an entity which identifies a location independently of the moment of time inside the time frame for which it is calculated (as opposed to the fingerprints that have been mentioned in Section 6.1.2 and which are extracted for each time bin)
- It is a list of N elements (where N is the number of APs which the user is associated with)
- Each element has the value 1 or 0
- If the element at position i in the list (where $i \in \{1..N\}$) is 1 it means that the associated AP (AP_i) has been found mostly with the value 1 in the fingerprints associated to the existing time bins (as presented in Section 6.1.2) and if it is 0 it means the opposite [5]

The way in which these signatures are create eliminates the problems that can appear in case interferences appear and disturb the presence of the signal from various APs for a limited amount of time. The location matching algorithm, in this case, can be as follows:

- We calculate the location signature for each location identified with our location extraction algorithm throughout a given number of days
- An entry (which is associated to a location) in the dictionary contains the location signature

[5] This means that if the signature of the location we are interested in has the element attributed to a given AP set to 1 then the AP has appeared in more time bins than the ones it was missing from during the given time frame. If it is 0, than the AP has been missing from more time bins than the ones that it was present in and that are associated to the give location.

- Before adding a new entry in the dictionary, we can first check the dictionary to see if any of the previous locations have a signature that is similar above a selected threshold to the one of the new location we are trying to add

- If a similar location is not found, we can proceed with adding a new entry in the dictionary for the new location

- If we find a location which resembles the new one in a proportion bigger than a given threshold then we can consider that the two are the same location

The similarity between two signatures is calculated by taking into consideration the APs that are set to 1 in either of the two signatures and the APs which are set to 0 but that are present in both signatures [6]. The similarity value si calculated as the number of APs that have the same value associated to them in both the signatures (either 0 or 1) and this number is then divided to the number of APs in the reunion.

The similarity result obtained for two signatures is compared to a given threshold in order to determine if the two signatures are referring to the same geographical location. We have experimented with threshold values between 65% - 98% and the results show that threshold values inside the interval 75% - 80% return the most accurate results for matching locations.

By analysing data for different users we found a case in which this method did not identify correctly that two locations where in fact the same one. The situation can be seen in Fig. 6.3.

As we can see, location L1 has 10 APs while location L2 has only 5. Since the APs in location L2 are also present in location L1 it is safe to assume that the remaining APs are not present in L2 due to technical problems or interferences. However, when calculating the signatures of the two locations and comparing them, by using a threshold value of 80% for considering them the same, we would only obtain that they are similar in a 50% proportion. In order to avoid this kind of situations, we have added to our algorithm another condition which goes as follows: if all the APs which have the value 1 in the signature of a location appear in the signature of an already found previous location, than the two can be considered the same location even if the similarity value is not

[6] The difference in APs in signatures is given by the fact that, when looking at data from different days, the APs which are scanned during these days might not always be the same. We are not keeping all the APs associated at any moment with a user because this leads to an unnecessary increase in the execution time.

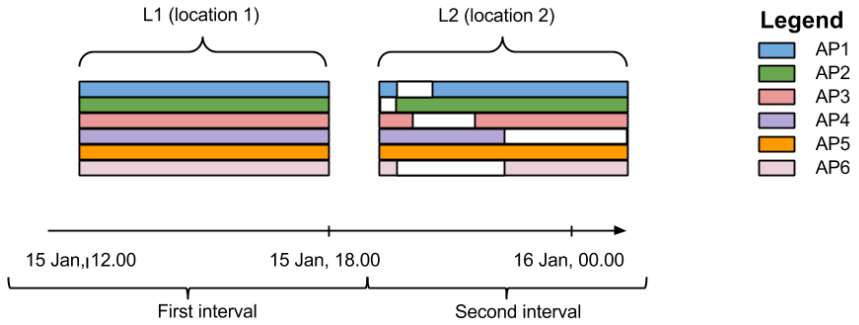

Figure 6.3: Example of locations which need to be matched despite their similarity being below the threshold

above the given threshold. This small improvement ensures that the algorithm performs very well even despite cases as this one.

This third possible solution performs very well for the test we have conducted over or data and as such it has been chosen for solving the matching problem for locations over different days.

Chapter 7

Entropy and predictability

The potential of using a real scientific approach in order to be able to predict where people will travel in the near or less near future can have an outstanding impact on the way in which the engineers design and construct infrastructures for cities, it can also impact the way in which we understand the transportation system and, not to mention, it could give us a new insight into how we can approach the solving of epidemics spreading [XL13] [DB08].

Data from SensibleDTU [SSS+14] allows us to explore for research purpose exactly how and why people move from a certain location to another. It gives us the opportunity to look more careful into our mobility patterns in order to try to understand how we can make use of these patterns to improve our world.

In order to explore the entropy and predictability of human mobility, we have conducted tests based on the data retrieved from a selection of users from the SensibleDTU database. We have selected 65 users from our original pool of 131 users in order to observe their movements throughout a period of 30 days. The reason behind discarding the remaining users was that their data was found to be missing important fields or they did not keep their mobile phones charged and open for the most part of the 30 days and thus their result could have jeopardize the study results. The current chapter focuses on presenting the results we have obtained as well as some observations regarding them.

7.1 Entropy

"Entropy is probably the most fundamental quantity capturing the degree of predictability characterizing a time series" [CS10]. Multiple studies ([RS14],[XL13], [MDX12], [CS10]) that aim at understanding the predictability of the human travel trajectories take into consideration different entropy measures which have different meaning and different levels of importance in correctly estimating the probability of choosing a location or another. The measures that are mentioned are the random entropy, the temporal uncorrelated entropy, the conditional entropy and the real entropy.

7.1.1 The random entropy

The formula for the random entropy of a random user i is given by

$$S_i^{rand} = log_2 N_i \tag{7.1}$$

Where N_i represents the number of unique locations that have been associated to the given user throughout the time frame that we are taking into consideration. This measurement can be used to reflect the predictability of the travel patterns of the given user in case we consider that each of the locations can be visited with the exact same probability.

We have calculated the random entropy by taking into consideration the locations visited by our selected users throughout a period of 30 days. Fig. 7.1 shows a histogram of the results which have been rounded after the second decimal. As we can see, most of the users have a random entropy of around 3.65 and the average random entropy is 3.87. This means that, in average, if a user was to choose randomly his or her next location, than they can be found in any of $2^{S^{rand}} = 2^{3.87}$ locations (which is approximately 14.62).

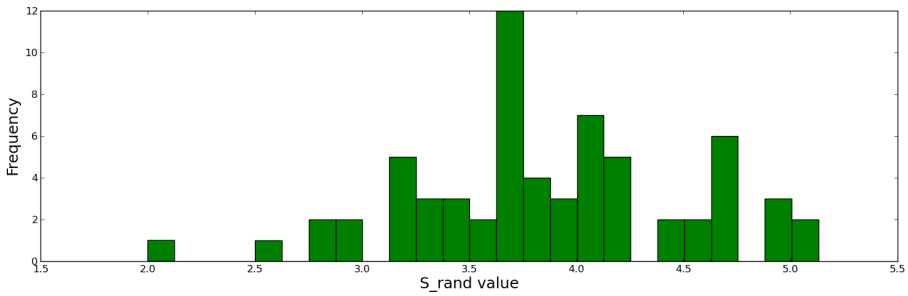

Figure 7.1: Histogram for S^{rand} values

7.1.2 The temporal uncorrelated entropy

The formula for the temporal uncorrelated entropy of a random user i is given by

$$S_i^{unc} = -\sum_{j=1}^{N_i} p_i(j) log_2[p_i(j)] \tag{7.2}$$

Where N_i represents the number of unique locations that have been associated to the given user throughout the time frame that we are taking into consideration and $p_i(j)$ represents the historical probability of the given user to visit location j. The present measurement incorporates the knowledge about what locations occur more often in the user's traveling patterns.

We have calculated the temporal uncorrelated entropy by taking into consideration the locations visited by our selected users throughout a period of 30 days. Fig. 7.2 shows the results which have been rounded after the second decimal. The average temporal uncorrelated entropy is 1.3. This means that, in average, if we are to base our supposition on the number of times each location has been visited in the past by a given user, then the user's next location can be found in any of $2^{S^{unc}} = 2^{1.3}$ locations (which is approximately 2.46).

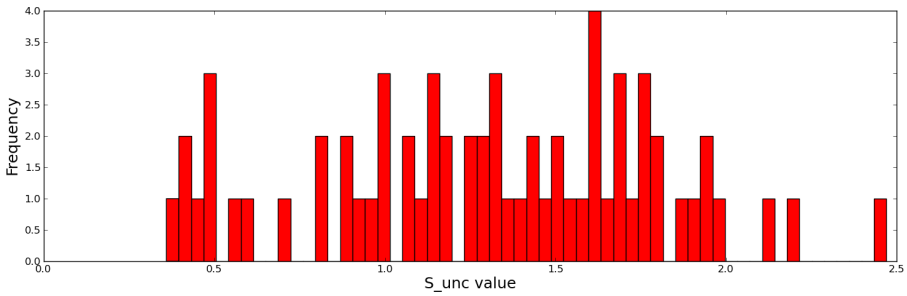

Figure 7.2: Histogram for S^{unc} values

7.1.3 The conditional entropy

The conditional entropy S_i^{cond} for a given user i is calculated based on the formula given in the paper [RS14]

$$S_i^{cond} = -\sum_{x_t \in X_i} \sum_{x_{t-1} \in X_i} p_i(x_{t-1}, x_t) log_2[p_i(x_t|x_{t-1})] \tag{7.3}$$

In this formula, x_t and x_{t-1} are possible locations, $p_i(x_{t-1}, x_t)$ is the probability of apparition of the subsequent locations x_{t-1} and x_t and $p_i(x_t|x_{t-1}) = p_i(x_{t-1}, x_t)/p(x_{t-1})$ represent the probability of the user being at location x_t at time t, considering that the previous location was x_{t-1}. The conditional entropy is equal to the temporal uncorrelated entropy in case we do not make any time correlations. Also it can be proved that $S^{cond} \leq S^{unc} \leq S^{rand}$ [CT06].

In [RS14] the authors introduce an extension for the conditional entropy in order to explore how the amount of previous knowledge affects the value of the conditional entropy. The extended formula is

$$S_i^{cond,k} = -\sum_{x_t \in X_i} .. \sum_{x_{t-k} \in X_i} p_i(x_{t-k}, .., x_t) log_2[p_i(x_t|x_{t-k}, .., x_{t-1})] \tag{7.4}$$

In this case, k represents the number of previous steps we know. With the present formula it can be observed that $S_i^{cond,0}$ is the same with S_i^{unc} and that $S_i^{cond,1}$ is the same as S_i^{cond}.

In Fig. 7.3 we have the representation for the conditional entropy calculated

considering that we have knowledge of a previous time window of $1 - 30$ time bins. It can be seen that the more information we have from the past, the value of the conditional entropy decreases.

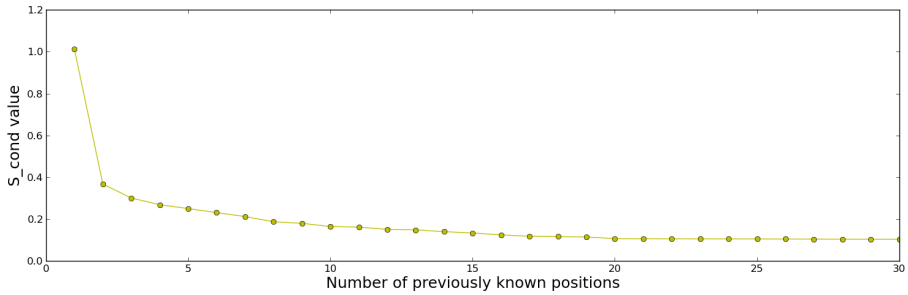

Figure 7.3: $S^{cond,k}$ for a given user, where $k \in \{1, ..30\}$

7.1.4 The real entropy

The real entropy of a user needs to be calculated by taking into consideration the different locations that the user has been to, the frequency with which he or she visits these locations, the time spent at the locations and the order in which the locations seem to follow through time. The relation between the actual entropy and the random and time uncorrelated ones is $S \leq S^{unc} \leq S^{rand}$.

If we consider that X_i represents the location of a user at time i, and h_n represents a sequence of n locations, then for a process $X = \{X_i\}$ the entropy can be written as

$$S \equiv \lim_{n\to\infty} \frac{1}{n} S(X_1, X_2, .., X_n) \tag{7.5}$$

$$= \lim_{n\to\infty} \sum_{i=1}^{n} S(X_i|h_{i-1}) \tag{7.6}$$

$$= \lim_{n\to\infty} \sum_{i=1}^{n} S(i) \tag{7.7}$$

Equation 7.5 is the definition given to entropy in [CT06], equation 7.6 reflects the application of the chain rule in the previous equation and $S(X_i|h_{i-1})$ represents the conditional entropy at step n in equation 7.7 [SQBB10].

The paper [CS10] presents another way in which the entropy for a user i can be written. If we consider that $T_i = X_1, X_2, \ldots, X_n$ represents the sequence of

locations which have been visited by user i and if we consider that $P(T_i')$ represents the probability of finding time ordered subsequence T_i' in the trajectory T_i, then the entropy of user i can be written as

$$S_i = -\sum_{T_i' \subset T_i} P(T_i') log_2[P(T_i')] \tag{7.8}$$

We have calculated a measurement of the entropy for our selected users and the distribution for the results which have been rounded after the second decimal can be seen in Fig. 7.4. The fact that the average value of the real entropy for the selected user is around 0.17 means that, in reality, the uncertainty about where a user will be traveling to next is $2^{0.17} = 1.12$ locations. These findings show that, the users we have been observing do not randomly choose their future locations, but in fact, their locations are well established by restrictions such as hours or days at which they need to be at work, or school as well as other patterns that govern our decisions most of the times.

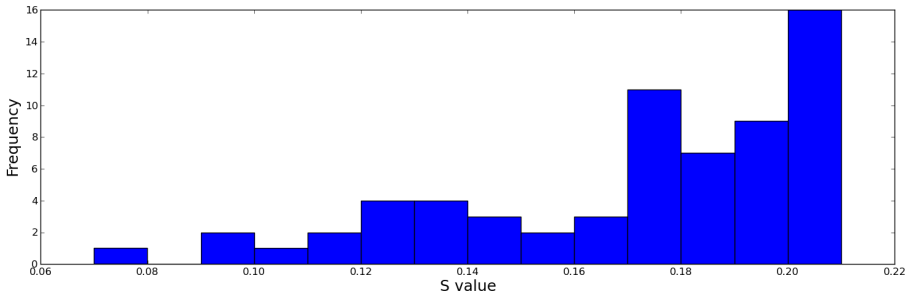

Figure 7.4: Histogram for S values

In Fig. 7.5 we have an overlay of the distribution for the random, temporal uncorrelated and real entropy.

7.2 Predictability

A measure that needs to be taken into consideration when discussing predictability is the probability (Π) of a predictive algorithm to correctly estimate the future locations a user will visit. According to the inequality of Fano [BOM08] [Tho62], a user for which the entropy is S and it has been calculated taking into consideration the fact that the user spends his or her time in either of N

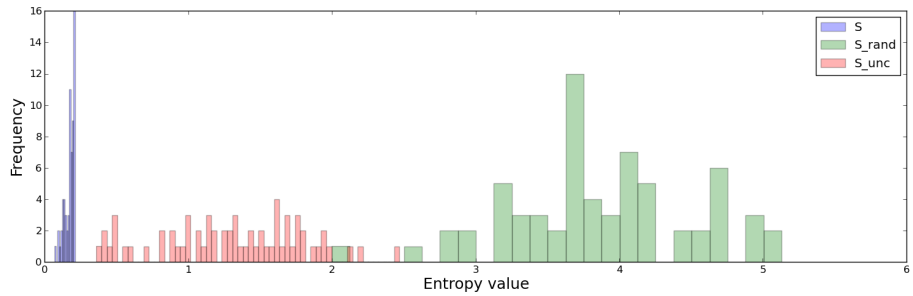

Figure 7.5: Histograms for S, S^{rand} and S^{unc} values

given locations, then his or her predictability (as it is presented in [CS10] and in [RS14] as well) is given by

$$\Pi \leq \Pi^{max}(S, N) \tag{7.9}$$

The value for Π^{max} can be determined from

$$S = H(\Pi^{max}) + (1 - \Pi^{max})log_2(N - 1) \tag{7.10}$$

by knowing that

$$H(\Pi^{max}) = -\Pi^{max}log_2(\Pi^{max}) - (1 - \Pi^{max})log_2(1 - \Pi^{max}) \tag{7.11}$$

By combining the two previous formula we obtain the equation

$$S + \Pi^{max}log_2(\Pi^{max}) + (1 - \Pi^{max})log_2(1 - \Pi^{max}) - (1 - \Pi^{max})log_2(N - 1) = 0 \tag{7.12}$$

One of the aims of our study is to estimate the predictability when it comes to the trajectory patterns of the selected users from the SensibleDTU database. After we have calculated their entropy and after knowing the number of locations each of them has been traveling in between for the duration of 30 days that we considered for our experiment, we only needed to calculate the entropy based on the mathematical formulae at 7.10 and 7.11. However, the equation

at 7.12, which would allow us to calculate the predictability for each user is a transcendental equation. This means that the equation can be solved either numerically or graphically.

In order to solve the equation we use a numerical method that helped us approximate the result. The method we use is the bisection method [BF85]. This method is applicable for solving an equation $f(x) = 0$ over a given interval $[a, b]$ for which the values $f(a)$ and $f(b)$ have opposing sings. In our case the f function is represented by the left side in the equation 7.12. Since the solution of our equation $f(x) = 0$ needs to be a number between 0 and 1 (the maximum predictability can only have a value in this interval), than we start the algorithm by verifying if by replacing the maximum predictability with 0 or 1 we have a solution or if the two values obtained this way have opposing signs. If they have opposing signs, we consider that $a = 0$ and $b = 1$. In case the function does not have opposing signs, we adjust the interval until we find either a solution or values for which the signs are opposing. The algorithm continues as following:

- As long as we do not exceed a previously set number of iterations we look for the middle of the interval given by the selected a and b numbers (let the middle be m)
- In case $f(m)$ is 0 or sufficiently close to 0 (we accept an error of order 10^{-3}), than we have found our solution, otherwise if the value of $f(m)$ has the same sign as $f(a)$ we move forward considering the interval $[m, b]$, if it has the same sign as $f(b)$ we continue by using the interval $[a, m]$
- We increment the number of iterations we have computed

The average predictability value for the selected users and through the given 30 days of observations is of approximately 98% which supports the observations made in previous studies (some of which have already been mentioned in the present paper, like, for example, [SQBB10], [MCG08] etc) regarding the fact that we seem to have a highly well established pattern of traveling that forms due to our daily habits.

Fig. 7.6 shows a histogram for the calculated predictability values for the selected users. In this particular case, the users for whom the values have been determined have things in common (for example they are all students at the Technical University of Denmark and they probably have similar ages) and they can be viewed as a community, yet even considering these similarities between the users, the results based on their data are interesting and seem to attest that, in general are mobility actions are highly predictable.

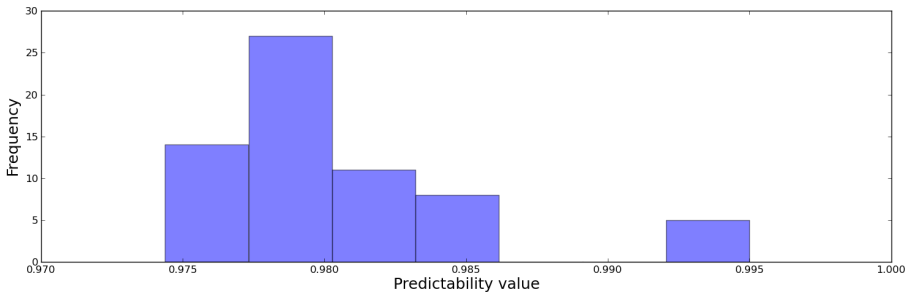

Figure 7.6: Histogram for predictability values

The difference of 5% between the value of the maximum predictability average we have obtained (98%) and the value obtained in the study presented in [SQBB10] (93%) can be caused by a multitude of factors. For example, the number of users which took part in the two studies is very different (65 users for the present study versus 45000 for the one presented by Song et.al.). There is also the difference between the types of data that have been analyzed. Song et. al. analyze data that comes from telephone towers while during our research we analyze Wifi data. The location for the users in our study can be closely monitored due to the fact that we have access to highly granular data, while in the case of the users observed by Song et. al., the data is collected only when the phone is used by the owners which can lead to periods of time when the location of a user is not known exactly. These are just a few examples of factors that can justify the difference between the results, however, what is worth considering is that, even though the experiments are quite different, the results seem to point in the same direction: our mobility patterns are rooted in our daily habits rather than being the result of spontaneity.

Chapter 8

An evaluation of Wifi positioning accuracy

Up to this point in our paper we have discussed the steps that have been taken in order to extract and analyze locations from Wifi data. The aim of the current chapter is to present the steps we have taken in order to evaluate the Wifi positioning in comparison to solutions that relay on GPS data.

The evolution of technology during the present days allows us to make use of devices that register and process high amounts of contextual information about their owners. Among these devices we can find our smartphones. Smartphones have become a commodity without which we can hardly imagine our day to day life. We use them for connecting and communicating with our friends, we use them for recreational activities such as "surfing" the web, or in order to keep ourselves up to date with the news in various domains that have captured our interest. However, aside from allowing us to access all these types of information while connecting us with the world around us, our smartphones have another ability which not so many of us completely understand. They can collect a large amount of contextual information about us. A strong example of how such information can be used in the interest of the mobile phone's owner is provided by the Google Now [GN] service. This service uses various contextual information that we make available through our Android phones in order to provide us with important information like traffic statistics for when we are

expected to travel, weather information for the destination we should be arriving at, or news about our favorite television show.

An example of contextual information that can be retrieved from our smartphones is our locations, or our traveling patterns. This is possible due to the GPS positioning that our smartphones allow. There are numerous studies which focus on the possibility of extracting human mobility patterns from the GPS data provided by mobile phones; among these studies we find [MGP10], [CLL14], [ZFL+07], [AS03].

The database from SensibleDTU (Section 3.1) contains, as it has been mentioned previously, numerous types of information among which we can also find GPS coordinates of the users that have accepted to be part of the scientific project. For the present study, we had the opportunity to employ part of this data in order to compare the results we obtain about the user locations from their Wifi data with possible stop locations provided from their GPS data.

8.1 Extracting stop locations from GPS data

There are different algorithm that can be used for extracting stop locations from GPS data. We present the three solutions that have been explored in [CLL14] and which we have taken into considerations as possibilities for extracting the stop locations.

8.1.1 Speed thresholding

This method relays on the calculation of speed in order to estimate if the data represents a stop location or a movement. The approximation of the speed between two given positions is calculated in this case as

$$Speed_i = \frac{Distance(pos_{i-1}, pos_i)}{timestamp_{pos_i} - timestamp_{pos_{i-1}}} \tag{8.1}$$

Based on the given dataset and the frequency of sampling, the calculated speeds can oscillate a lot. In [CLL14] the proposed solution for this problem is to have the data grouped into time bins of a given size and to consider that the position which corresponds to each time bin is given by the median of the samples in the

bin, thus allowing the calculation of the speed between the bins instead of the samples.

After the speeds are calculated, the samples for which a high speed [1] is identified can be discarded as they can represent movement. The remaining samples can be, therefor, associated to stop locations and can be grouped into different stop locations.

8.1.2 Gaussian Mixtures Model

A different approach presented in [CLL14] for grouping samples into stop locations is to observe the overall distribution of the existing samples without taking into consideration the timestamps. The idea is to observe the clusters that form and to consider as stop locations the clusters of samples which have a higher density.

The clusters are created by using a Gaussian Mixtures Model that attributes samples to different clusters which are modeled as Gaussian distributions which have unknown parameters. After a sample is allocated to a cluster, consecutive samples that have been assigned to the same cluster can be grouped and consider to form stop locations. The size of the clusters are determined by a given parameter.

8.1.3 Distance grouping

The main idea behind the distance grouping algorithm, as it has been presented in [CLL14], is that a sequence of location samples which seem to be geographically close to each other [2] can be grouped into a stop location.

The stop locations are created as follows:

- The samples are taken into consideration in the ascending order of their timestamps
- A stop location has initially only one location sample L_i attributed to it

[1] A speed is considered to be high if it is above a defined $Speed_{max}$ value

[2] The location samples have been identified within a given maximum distance D_{max} of each other

- All the following k samples as long as all samples $j \in \{i+1, k\}$ respect the condition $Distance(L_j, L_i) < D_{max}$ are added to the location
- The process of constructing the next stop location is started again from location L_{i+k+1}

8.2 Comparing results obtained with distance grouping algorithm

For the purpose of analysing the locations we have obtained from the Wifi data we have implemented the *distance grouping algorithm*. Our implementation considers that the maximum distance D_{max} that determines the way in which GPS samples are grouped is of 60 m. The way in which this value is selected affects the results as a very large large value for D_{max} leads to the merging of different stop locations while a very small value leads to stop locations being unnecessarily divided into more smaller stop locations. The 60 m value has proved to be a good approximation as it has been determined empirically from the data that it provides acceptable results.

We also try to be consistent with the fact that for the Wifi data we consider that a user needs to be close to a given geographical location for at least 5 minutes in order for it to be considered a stop location and not a transition location. In order to keep this assumption for the locations extracted from the GPS data, we discard the stop locations for which the user does not seem to be spending at least 5 minutes within their limits. To calculate the time the user spends in a GPS identified stop location we calculate the difference between the timestamps of the first sample that is associated with the position and the last sample that is still considered a part of the stop location. If the difference is equal or above 5 minutes, the stop location is kept.

Fig. 8.1 represents a quantitative measurement for the number of locations that a group of 65 users change during a period of 30 days according to the Wifi data, while Fig. 8.2 represent a quantitative measurement for the number of locations the same group of users changes during the same period of time according to the GPS data. As we can see, by using the GPS data we can identify a biger number of locations for a given time frame than by using the Wifi data. The average number of locations changed during a month per user according to the GPS data is approximately 153, while the average number identified for Wifi data is approximately 80. Considering these results we wanted to take a look at a time frame from the perspective of the locations identified with both Wifi and GPS data in order to understand the differences between the estimations.

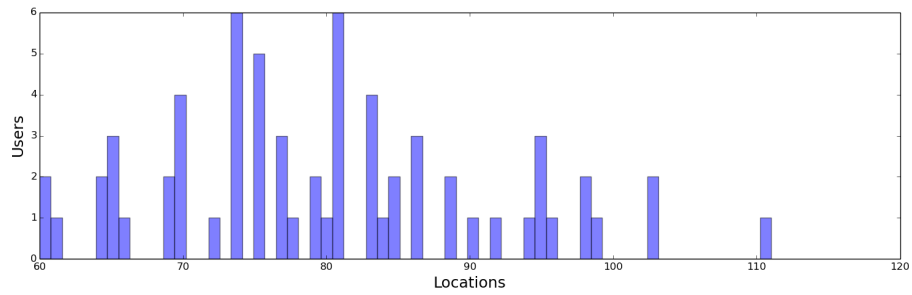

Figure 8.1: Count of location changes done during 30 days by 65 users according to Wifi data

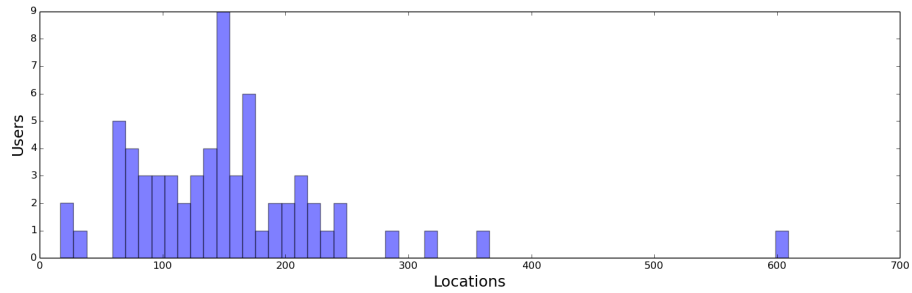

Figure 8.2: Count of location changes done during 30 days by 65 users according to GPS data

In Fig. 8.3 we can see the predominant APs that have been scanned during a 2 days time frame for an user. Fig. 8.4 displays the locations identified considering the Wifi data (first line of colors) and the locations extracted from the GPS data (second line of colors).

From these images we can observe that the GPS data has, indeed, a better accuracy when dealing with locations. For example, from the GPS data we can identify that from Thursday morning and up to Thursday at noon, the user is recorder to have changed location three times. In the Wifi data, only two locations are identified withing the same time frame. A reason why this can be possible is that the stop locations identified from GPS data are grouped together if they are less than 60 meters apart, while the Wifi stop locations are calculated based on the visibility of the Wifi APs. In this case, if some APs have a range of visibility which extends 60 m, stop locations which can be identified as being different based on GPS can be considered the same based on Wifi data.

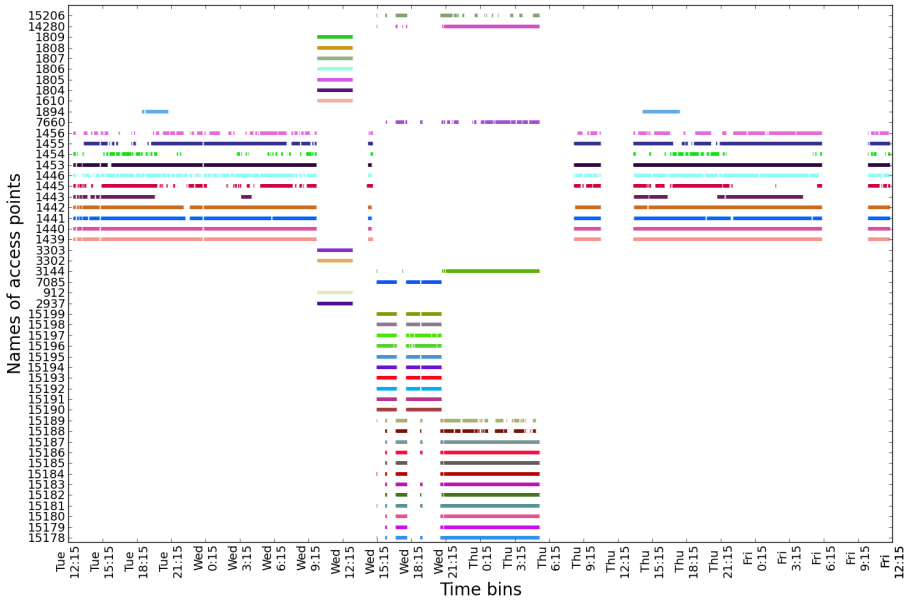

Figure 8.3: The most common 50 APs for userX during 3 days scan records

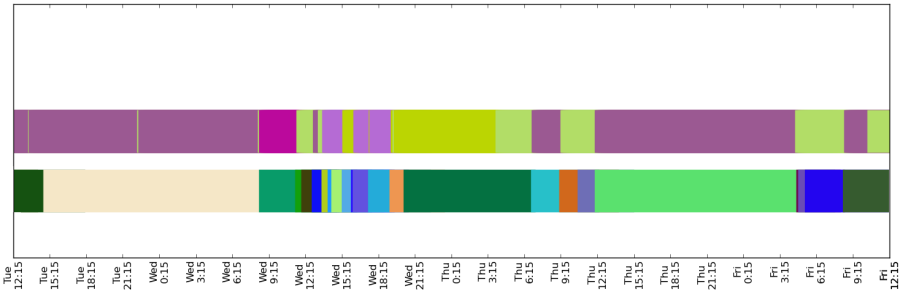

Figure 8.4: Wifi and GPS locations identified for userX throughout 3 days

Despite some differences, the two figures do present similarities and are clearly associated to the same time frame and the same user (for example, the time between Thursday at around 12.15 until Friday at around 4.00 is identified in both data as one stop location). However, the need for further work in order to improve the way in which we estimate locations based on Wifi data is needed. In this sense, an analysis of stop durations for both Wifi and GPS data and an analysis of the fraction of stops that approximately overlap can be a good starting point for further research.

Chapter 9

Discussion and future work

During the work conducted for our study we have come across a number of interesting findings and observations. This chapter contains a summation of these surfacing results as well as a list of possible directions and subtopics that can be explored during future research on the subject.

In order to capture results that can reflect general facts about the way human mobility patterns emerge, it is important for the conducted studies to incorporate the usage of a high amount of data. Despite the quantity of data being relevant for the results, the importance of the quality of the data can exceed it. Considering this, a thorough cleaning of the Wifi data proved to be of a high importance for our present research. By eliminating the noise and the excessive signals from various sources that were not of interest for the present work, we have obtained data that allowed us to make accurate observations on human mobility.

Despite the data clean up, it has been easy to notice from the very beginning that the signal strength received from the high number of access points does not tend to stay constant even if the location seems to be maintained by the user. Interferences influence the signal strength [MCWA10] and can complicate the algorithm used for identifying locations based both on access points BSSIDs and RSS values. We have analyzed the possibility of using different average signal calculations (Sections 5.2.4 and 5.2.5) in order to smooth out the spikes caused

by interferences. The averaging, especially the running average, did improve the data by removing a part of the spikes, however this proved to not be sufficient for the present case. Previous research projects show that an approach that does not necessarily require the use of signal strength for the visible access points, but just the knowledge of the signals being present or not in a given point can be a good starting point for determining a location [LJ09]. We have, therefor, determined the presence of the various access points across different time frames and we have used the results in order to explore the possibility of determining locations based on this information.

It is worth mentioning that additional work can be done for improving our understanding of interferences and noise that can cause problems when working with Wifi data. The present paper presented a few techniques which have been used in order to eliminate possible noise, however, future improvements can be added for obtaining data which has an even higher qualitative value.

In order to extract locations from Wifi data we have worked with three different algorithms: based on networks, k-means clustering and based on the use of Hidden Markov Models. The algorithm that uses the construction of a network based on the presence of the access points at given times did not lead to acceptable results for the duration of our study. However, the k-means clustering and the algorithm based on Hidden Markov Models combined with a k-fold cross validation algorithm in order to estimate the correct number of locations that were to be expected for a user lead to good approximations for stop locations. For the data we have used, the algorithm based on Hidden Markov Models, has statistically given better approximations, but further research is needed in order to determine if either of the two methods is better than the other or they just behave better in various circumstances. Another thing which is worth taking into consideration for further research is that, even though during our study, we did not reach acceptable results by the use of the network method, further research is needed in order to determine if the idea behind this approach can produce note worthy findings. The benefits are that, by identifying clear characteristics of the locations, this enables the construction of an algorithm that can have a lower execution time for extracting locations from large amounts of data. Also, an additional interesting focus point can be represented by researching the implementation of an improved method that can be used for estimating the expected number of locations for a given time frame. The present paper has presented the use of cross validation as a method of determining the best possible estimation, however, Hidden Markov Models constitute a highly effective model which, possibly, with further research can be able to eliminate the use of estimation and allow the exact determination of the number of locations which can be observed during a given sequence of data. Other options can be explored and tested, as well, for achieving the best results.

Due to the using of cross validation in order to determine the approximate number of locations we are to expect for each user during a given time frame, as well as the complexity of the location determination algorithms, the execution time rises exponentially if the amount of input data is high. In order to avoid this problem we have determined locations over short one day intervals. The locations estimated for different days needed to be compared and analyzed in order to determine when they coincided. This offered us the opportunity to explore different methods in order to determine the location matching (Chapter 6). We have also compared our extracted locations with locations extracted from GPS data and the results support that, even though the GPS data offers better location estimations, the Wifi data can provide acceptable results as well. Even though we have taken into consideration various possibilities that can be used in order to determine if two locations which have been identified in different iterations by a location extraction algorithm can, in fact, be considered to be the same location, further work can be done on this subject. Even though the estimations which the present solution gives are accurate most of the times, additional improvements can be made and further research can lead to the development of a new method that can exceed the benefits of the one that has been used in this case.

The locations extracted from a selection of 65 of users from our original pool of 131 users has allowed us to explore the predictability of human travel trajectories. The results we have reached support previous studies ([CS10], [MCG08], [RS14], [DB06] etc.) which indicate that the mobility patterns we display seem to have a high degree of predictability. This can prove that, in fact, we could be less spontaneous than we believe we are and that our behaviour is deeply rooted in habits. However, further research can still be done and is needed for estimating the predictability of human mobility. The results we have obtained are based on the data that has been collected through a period of one month and is provided by a focus group of 65 people who have in common the fact that they are students at the Technical University of Denmark. Further research on a larger group of people possibly with various backgrounds could give additional interesting results and would be needed in order to make definitive affirmations about emerging mobility and predictability patterns.

During our work we have estimated the accuracy of our algorithms and proposed solution by using visualization of the results. Future work can be done in order to employ the use of various tools that can be used to make exact calculations for the accuracy of the proposed solutions.

During our work we have debated what is a good approximation for the time an user can be expected to stay at a given location in order to consider it a stop location rather than a transit location. We have proposed the use of a 5 minute time frame, however further studies can be conducted in order to estimate what

is the average time a person spends at a given location based on the location type. This can lead to further considerations and adaptations for algorithms that try to extract stop locations both from Wifi data, but also from other types of data, like ,for example, GPS data.

Clearly the path of exploring the subject of human mobility and predictability of human travel trajectories is far from reaching an end. Numerous studies, the present one included, have been conducted on this topic, however there are still an enormous number of questions that need answers and further work is needed in order to be able to fully understand what determines our trajectories and how we can use the knowledge in order to increase our quality of life. The topic continues to remain of high interest as the possible results have an applicability potential that expands over a variety of fields and domains such as the prevention of the spreading of epidemics, the design of better urban and transportation infrastructures, and many other areas of interest.

Chapter 10

Conclusions

This paper discusses the steps which have been taken in order to study the inferring of human mobility patterns from Wifi data. The focus of the study has been divided in between different areas of this subject.

We have firstly analyzed what defines a Wifi determined location. We have taken into consideration different approaches that can be used in order to extract locations. We considered determining locations based on access points' BSSIDs and RSS. This option has led to us trying to find a solution for a better noise elimination. We have experienced with different averages that aimed to smooth the spikes in the signal strength, however we have come to the conclusion that for the used data a better approach is to only take into consideration the BSSIDs of the access point and the knowledge if an access point is present or not in a given time bin. The information extracted by analyzing when the access points are visible has been used in order to determine stop locations.

The extraction of the locations based on the identity and presence of various access points throughout different time frames has proved to be a challenging and complex task. We have tested three different methods: an algorithm based on the usage of networks, an algorithm based on k-means clustering and an algorithm based on Hidden Markov Models. The first algorithm did not lead to satisfactory results in the present study. Between the second two approaches, the algorithm based on Hidden Markov Models has statistically had better re-

sults for our used data, however further research is needed in order to make a definitive assumption on which of the two algorithms can prove to be better to use in different circumstances. Also, in order to use either of k-means or Hidden Markov Models based algorithms we needed to first determine the number of locations we were expecting the algorithms to identify. This has been achieved by using cross validation in order to rank the results based on different numbers of expected locations.

In order to conduct the study over a large amount of data we needed to run the location identification algorithm on data collected during smaller time frames (e.g. one day). The resulting locations have been compared to each other. This was needed in order to determine if a location appeared in the results from a different iteration. We have also compared the resulting locations extracted from Wifi data to locations which have been extracted from analogue GPS data. The results have been satisfactory in the sense that, even though, as expected, GPS data can offer more detailed results about location changes, the accuracy of the Wifi results is not to be overlooked.

By knowing the different locations and the sequence in which they occurred throughout time we were able to calculate various entropy values for a selected pool of users. We have also been able to determine that the users which have been selected to be part of the present study present a high degree of predictability as far as human mobility is concerned. This observation supports previous results which have been made regarding this topic, yet further research on an even bigger data set and longer period of time might be needed in order to establish if this particular observation can be considered a fact.

We have finalized the paper by making a summary of the most important observations that have emerged throughout our work as well as a series of suggestions about possible future areas of interest for the present topic.

To conclude, we would like to state that the work in the present field is far from being complete. There are many questions that still require answers and many opportunities for improvement and we feel that the findings presented in the current paper, as well as previous works can constitute a solid ground for further research projects. The topic in itself is of a high interest for the future as any results can be used to drastically improve the quality of life for generations to come.

Appendix A

Appendix

A.1 Variations for signal strength visualization over time

This section contains various visualizations for different users' scanned access points over time. On the x axis we have the time frame, while on the y axis we have the signal strength for the identified access points. The legend presents only the top 10 predominant access points (which have appeared the most during scans), however the plot displays all access points. The figures are Fig. A.1, Fig.A.2, Fig. A.3, Fig.A.4.

A.2 Sample density for APs identified for a user

This section contains the visualization for the signal strength of different APs that have been identified as being associated to a user throughout a period of 1 day (Fig. A.5) as well as the sample density visualizations for the top various APs that were scanned throughout this time (Fig. A.6,Fig. A.7,Fig. A.8).

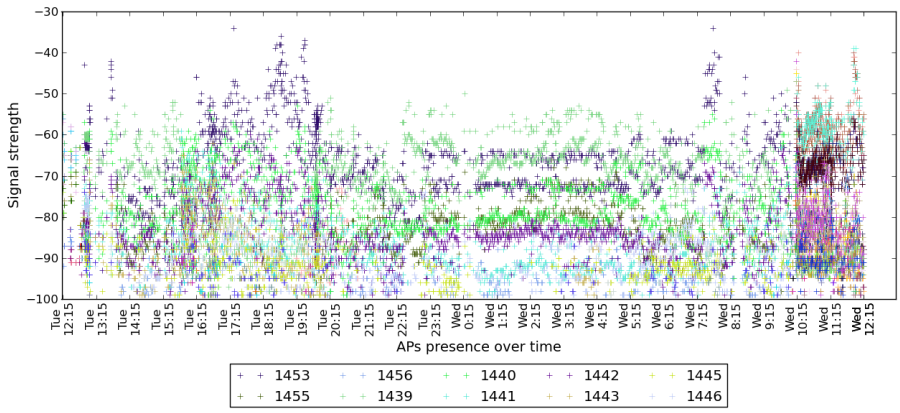

Figure A.1: Example of the APs registered for userX throughout one day with "+" markers

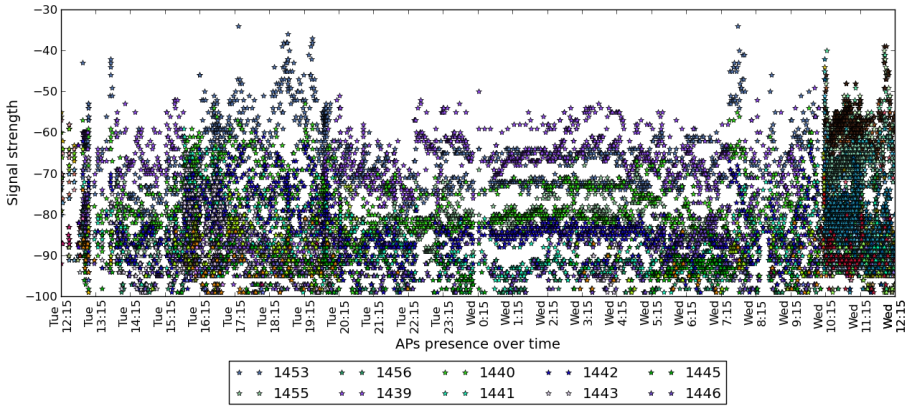

Figure A.2: Example of the APs registered for userX throughout one day with "*" markers

A.2.1 Average signal strength for APs identified for a user

This section contains the visualization for the average signal strength of APs 15188 (Fig. A.9), 15190 (Fig. A.10) and 3144 (Fig. A.11) calculated for 5 minutes time bins over the course of one day from the data gathered for a given user.

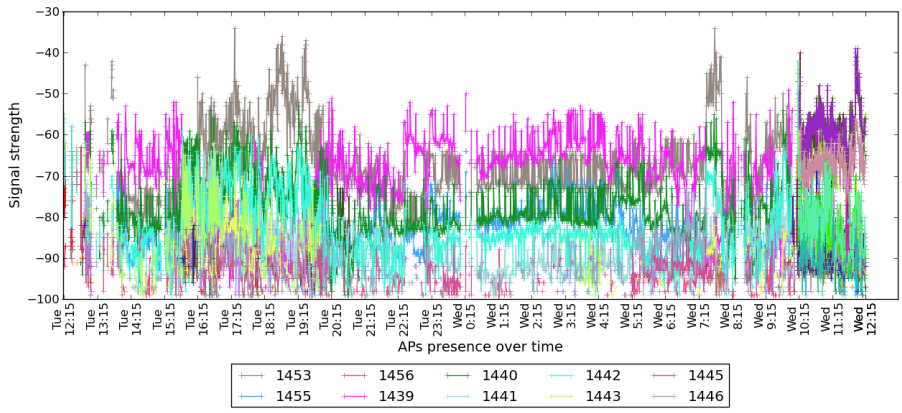

Figure A.3: Example of the APs registered for userX throughout one day with "+" and line markers

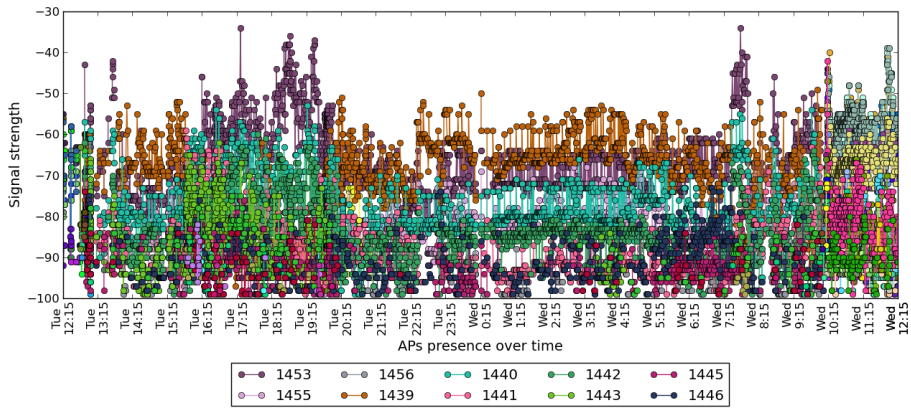

Figure A.4: Example of the APs registered for an user throughout one day with "o" and line markers

A.2.2 Running average signal strength

This section contains the visualization for the running averages calculated for 2 (Fig. A.13), 5 (Fig. A.14) and 10 (Fig. A.15) minutes time bins for AP 1613 identified in a time frame of one day (Fig.A.12).

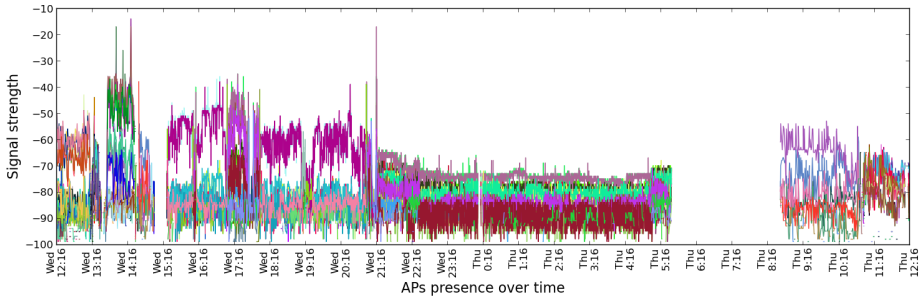

Figure A.5: Example of the APs registered for userX throughout day 2

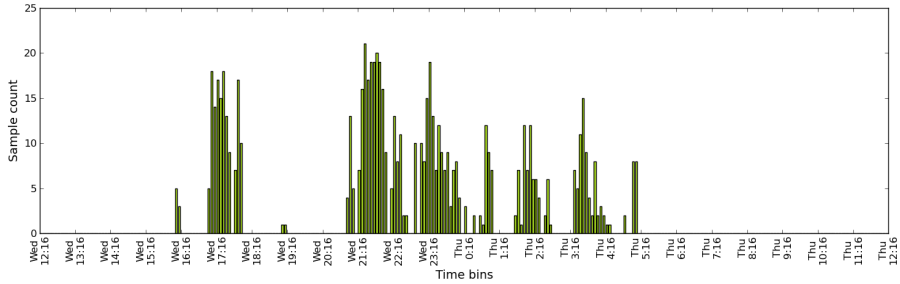

Figure A.6: Sample density of AP 15188 for userX

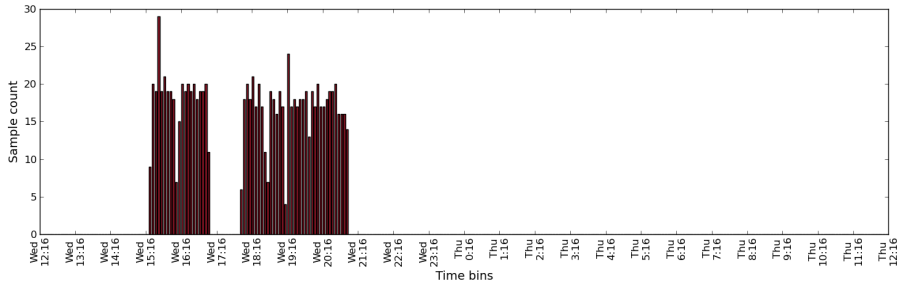

Figure A.7: Sample density of AP 15190 for userX

A.2.3 Signal presence

This section contains the visualization for the presence of APs for a period of 2 days for an user from the SensibleDTU database (Fig. A.17). The presence for APs is determined for 5 minutes time bins over the 2 days. Fig. A.16 presents

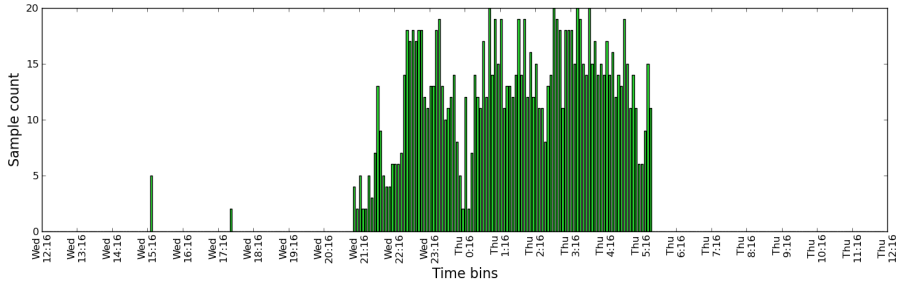

Figure A.8: Sample density of AP 3144 for userX

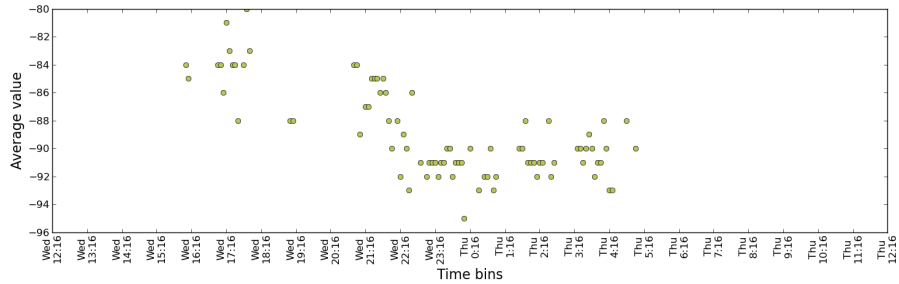

Figure A.9: Average signal strength of AP 15188 for an user

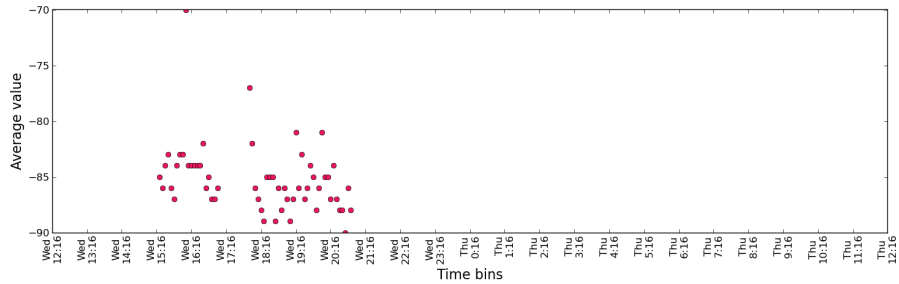

Figure A.10: Average signal strength of AP 15190 for an user

all the APs (and their signals) visualized for the same 2 days.

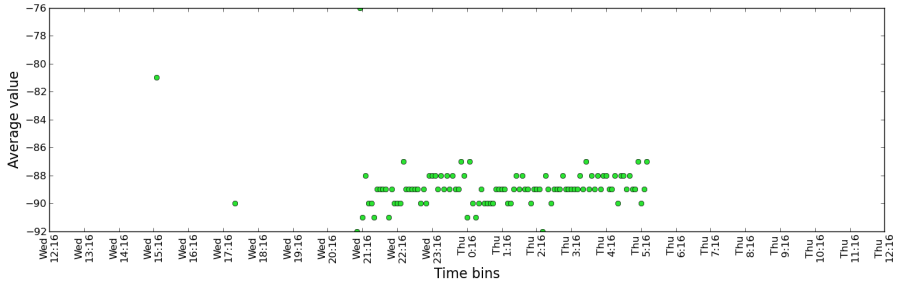

Figure A.11: Average signal strength of AP 3144

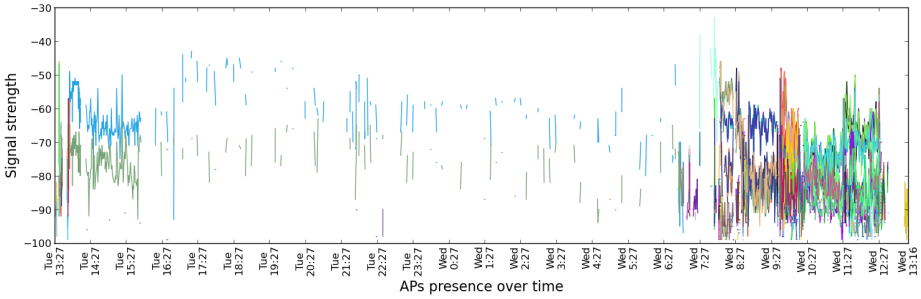

Figure A.12: Example of APs presence over time for userT

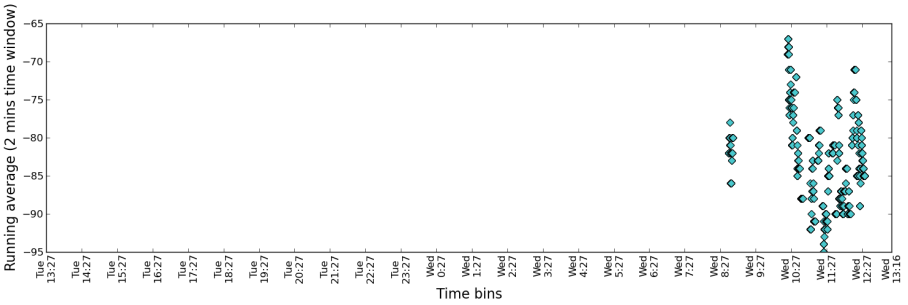

Figure A.13: Running average for AP 1613 for userT during 1 day (2 minute time bins)

A.2.4 Locations extracted using k-means

An example of locations that have been identified by using the k-means algorithm over a period of 3 days for a given user can be seen in Fig. A.19. Fig. A.18

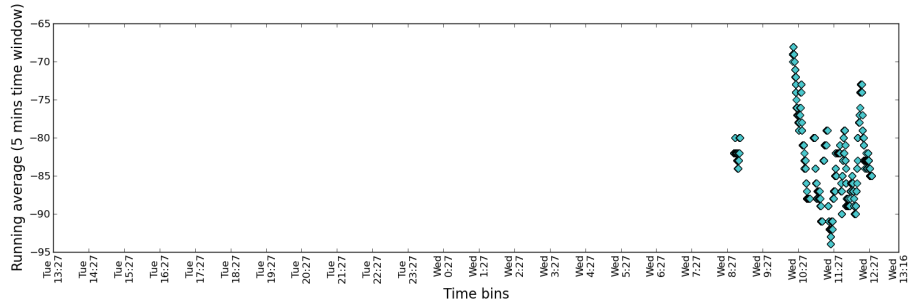

Figure A.14: Running average for AP 1613 for userT during 1 day (5 minute time bins)

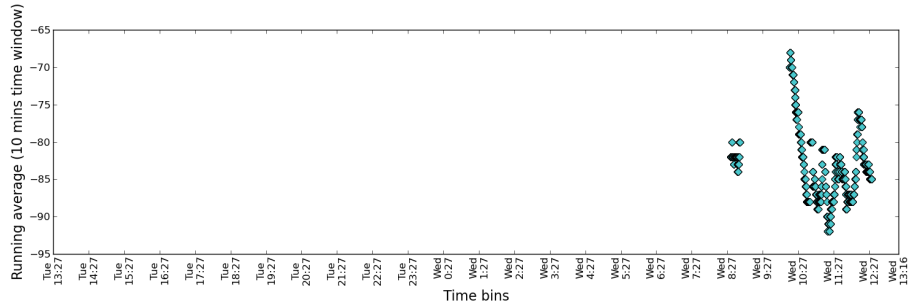

Figure A.15: Running average for AP 1613 for userT during 1 day (10 minute time bins)

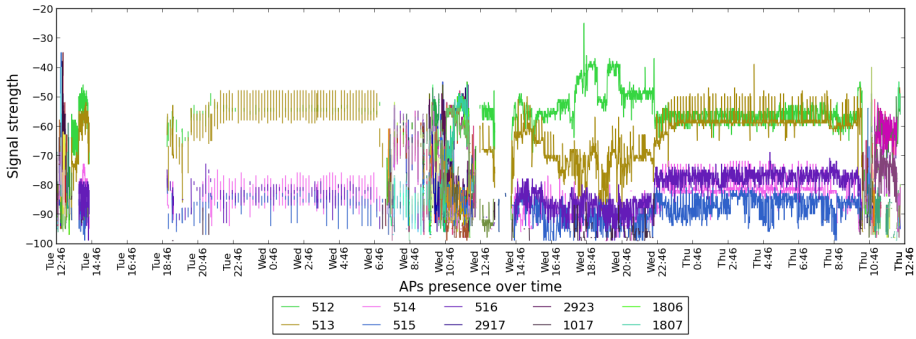

Figure A.16: Scanned APs for an user throughout a duration of 2 days

represents the presence of the most common 50 APs which have been scanned throughout the same time for the same user.

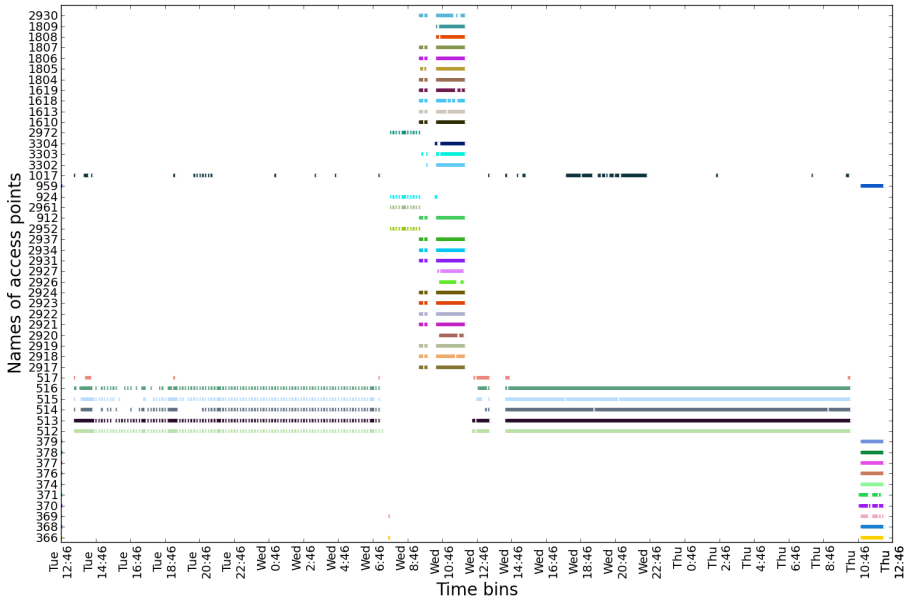

Figure A.17: The most common 50 APs for an user during 2 days (presence visualization calculated for 5 minutes time bins)

A.2.5 Locations extracted using HMM

An example of locations that have been found for a given user throughout 3 days can be seen in Fig. A.21. They can be easily mapped to the locations we can observe by looking at Fig. A.20 where we have the presence of the APs which have been scanned throughout the same amount of time for the same user.

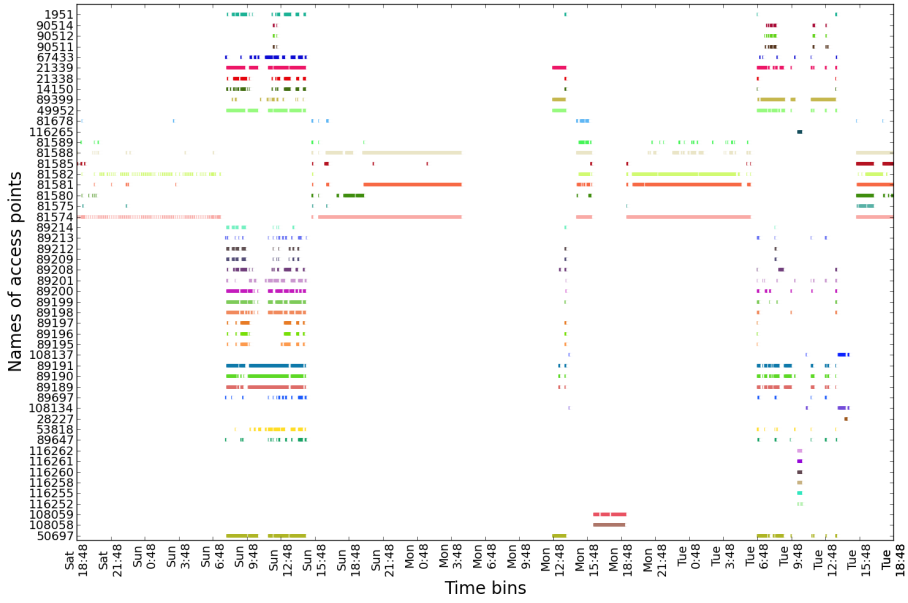

Figure A.18: The most common 50 APs for an user during 3 days (presence visualization calculated for 5 minutes time bins)

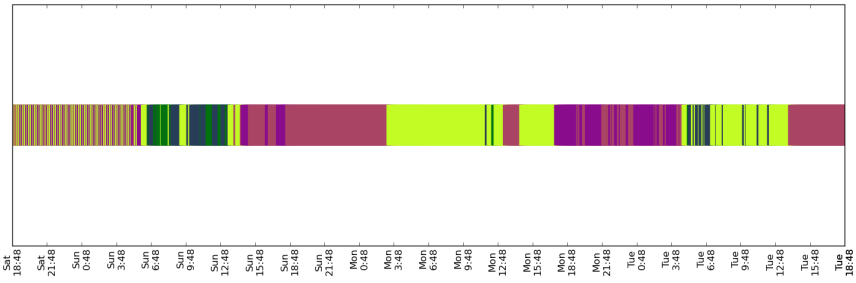

Figure A.19: Locations estimated with k-means for an user for 3 day

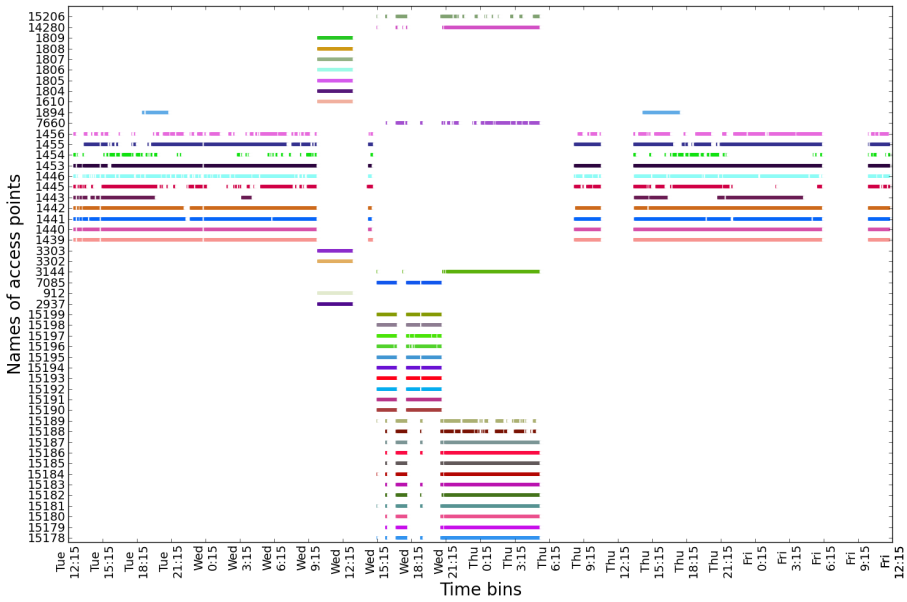

Figure A.20: The most common 50 APs for an user during 3 days (presence visualization calculated for 5 minutes time bins)

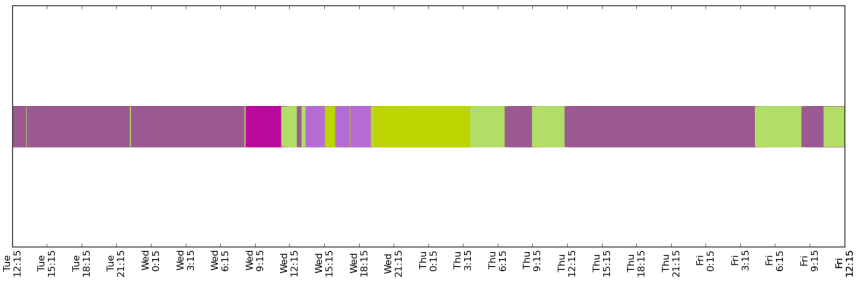

Figure A.21: Locations estimated with HMM for an user for 3 day

Bibliography

[AGB13] Fereshteh Asgari, Vincent Gauthier, and Monique Becker. A survey on human mobility and its applications. *CoRR*, abs/1307.0814, 2013.

[AGGP09] Spiros Athanasiou, Panos Georgantas, George Gerakakis, and Dieter Pfoser. Utilizing wireless positioning as a tracking data source. In *Advances in Spatial and Temporal Databases*, pages 171–188. Springer, 2009.

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

[AS07] Sherif Akoush and Ahmed Sameh. Mobile user movement prediction using bayesian learning for neural networks. In *Proceedings of the 2007 international conference on Wireless communications and mobile computing*, pages 191–196. ACM, 2007.

[AS14] Alex Pentland David Lazer Sune Lehmann Arkadiusz Stopczynski, Riccardo Pietri. Privacy in sensor-driven human data collection: A guide for practitioners. *CoRR*, abs/1403.5299, 2014.

[ASA10] F Alsehly, Z Sevak, and T Arslan. Improving indoor positioning accuracy through a wi-fi handover algorithm. In *Proceedings of the 2010 International Technical Meeting of the Institute of Navigation - ITM 2010*, 2010.

[Bal03] P. Ball. The physical modelling of human social systems, 2003.

[BF85] Richard L. Burden and J. Douglas Faires. *Numerical Analysis*, chapter 2. PWS Publishers, 3 edition, 1985.

[BOM08] Anthony Brabazon, Michael O'Neill, and Dietmar Maringer. Natural computing in computational finance, 2008.

[BP00] Paramvir Bahl and Venkata N Padmanabhan. Radar: An in-building rf-based user location and tracking system. In *INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, volume 2, pages 775–784. IEEE, 2000.

[CCJ10] Xin Cao, Gao Cong, and Christian S. Jensen. Mining significant semantic locations from gps data. *Proc. VLDB Endow.*, 3(1-2):1009–1020, September 2010.

[CCLK05] Yu-Chung Cheng, Yatin Chawathe, Anthony LaMarca, and John Krumm. Accuracy characterization for metropolitan-scale wi-fi localization. In *Proceedings of the 3rd International Conference on Mobile Systems, Applications, and Services*, MobiSys '05, pages 233–245. ACM, 2005.

[CJK05] Jernej Copic, Matthew O. Jackson, and Alan Kirman. Identifying community structures from network data via maximum likelihood methods, 2005.

[CLL14] Andrea Cuttone, Sune Lehmann, and Jakob Eg Larsen. Inferring human mobility from sparse low accuracy mobile sensing data. In *3rd ACM Workshop on Mobile Systems for Computational Social Science(MCSS 2014)*. ACM, 2014.

[CML11] Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.

[CS10] 1 2 3 Nicholas Blumm 1 2 Albert-László Barabási 1 2 * Chaoming Song, Zehui Qu. Limits of predictability in human mobility. *Science*, 327, 2010.

[CSC+06] Mike Y. Chen, Timothy Sohn, Dmitri Chmelev, Dirk Haehnel, Jeffrey Hightower, Jeff Hughes, Anthony LaMarca, Fred Potter, Ian Smith, and Alex Varshavsky. Practical metropolitan-scale positioning for gsm phones. In *Proceedings of the 8th International Conference on Ubiquitous Computing*, UbiComp'06, pages 225–242. Springer-Verlag, 2006.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[DB06] T. Geisel D. Brockmann, L. Hufnagel. The scaling laws of human travel. *Nature*, 439, 2006.

[DB08] F. Theis D. Brockmann. Money circulation, trackable items, and the emergence of universal human mobility patterns. *Pervasive Computing, IEEE*, 7, 2008.

[Dra67] Alvin W. Drake. Fundamentals of applied probability theory, 1967.

[Edd04] Sean R Eddy. What is a hidden markov model? *Nat Biotech*, 22(10), 2004.

[EKHH13] Reda A El-Khoribi, Haitham S Hamza, and MA Hammad. Indoor localization and tracking using posterior state distribution of hidden markov model. In *Communications and Networking in China (CHINACOM), 2013 8th International ICST Conference on*, pages 557–562. IEEE, 2013.

[FBSW08] Bo Fu, Gábor Bernath, Ben Steichen, and Stefan Weber. Wireless background noise in the wi-fi spectrum. In *Wireless Communications, Networking and Mobile Computing, 2008. WiCOM'08. 4th International Conference on*, pages 1–7. IEEE, 2008.

[Fli03] Rob Flickenger. *Building Wireless Community Community Networks*. O'Reilly & Associates, Inc., Sebastopol, CA, USA, 2 edition, 2003.

[GCP+06] Y. Guo, P. Corke, G. Poulton, T. Wark, G. Bishop-Hurley, and D. Swain. Animal behaviour understanding using wireless sensor networks. In *Local Computer Networks, Proceedings 2006 31st IEEE Conference on*, pages 607–614, Nov 2006.

[Gep] Gephi - the open graph viz platform.

[GL96] G. Maguire Jr G. Liu. A class of mobile motion prediction algorithms for wireless mobile computing and communication. *Mobile Networks and Applications*, 1, 1996.

[GN] Google now.

[Hyn09] Rob J Hyndman. Moving averages, 2009.

[IHO13] Yusuke Inatomi, Jihoon Hong, and Tomoaki Ohtsuki. Hidden markov model based localization using array antenna. *International journal of wireless information networks*, 20(4):246–255, 2013.

[JP04] W. Trumler T. Ungerer L. Vintan J. Petzold, F. Bagci. Global state context prediction techniques applied to a smart office building, 2004.

[JS99] Oliver P John and Sanjay Srivastava. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(1999):102–138, 1999.

[Kal60] R. E. Kalman. A new approach to linear filtering and prediction problems, 1960.

[KC11] Jahyoung Koo and Hojung Cha. Autonomous construction of a wifi access point map using multidimensional scaling. In *Pervasive Computing*, pages 115–132. Springer, 2011.

[Koh95] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. pages 1137–1143. Morgan Kaufmann, 1995.

[LBG80] Yoseph Linde, Andres Buzo, and Robert M Gray. An algorithm for vector quantizer design. *Communications, IEEE Transactions on*, 28(1):84–95, 1980.

[LHK+09] Kyunghan Lee, Seongik Hong, Seong Joon Kim, Injong Rhee, and Song Chong. Slaw: A new mobility model for human walks. In *INFOCOM 2009, IEEE*, pages 855–863, April 2009.

[LJ09] Jakob Eg Larsen and Kristian Jensen. Mobile context toolbox: An extensible context framework for s60 mobile phones. In *Proceedings of the 4th European Conference on Smart Sensing and Context*, EuroSSC'09, pages 193–206. Springer-Verlag, 2009.

[Llo06] S. Lloyd. Least squares quantization in pcm. *IEEE Trans. Inf. Theor.*, 28(2):129–137, September 2006.

[Mac67] J. Macqueen. Some methods for classification and analysis of multivariate observations. In *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[MCG08] A. L. Barabasi M. C. Gonzalez, C. A. Hidalgo. Understanding individual human mobility patterns. *Nature*, 453, 2008.

[MCWA10] Aniket Mahanti, Niklas Carlsson, Carey L. Williamson, and Martin F. Arlitt. Ambient interference effects in wi-fi networks. In *Networking*, volume 6091 of *Lecture Notes in Computer Science*, pages 160–173. Springer, 2010.

[MDA05] Geoffrey McLachlan, Kim-Anh Do, and Christophe Ambroise. *Analyzing microarray gene expression data*, volume 422. John Wiley & Sons, 2005.

[MDX12] R-C Marin, Ciprian Dobre, and Fatos Xhafa. Exploring predictability in mobile interaction. In *Emerging Intelligent Data and Web Technologies (EIDWT), 2012 Third International Conference on*, pages 133–139. IEEE, 2012.

[MGP10] Raul Montoliu and Daniel Gatica-Perez. Discovering human places of interest from multimodal mobile phone data. In *Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia*, MUM '10, pages 12:1–12:10. ACM, 2010.

[MK06] S. Kim M. Kim, D. Kotz. Extracting a mobility model from real user traces. *The IEEE INFOCOM Proceedings*, 2006.

[MM07] Mirco Musolesi and Cecilia Mascolo. Designing mobility models based on social network theory. *SIGMOBILE Mob. Comput. Commun. Rev.*, 11(3):59–70, 2007.

[MNRS07] Carlo Morelli, Monica Nicoli, Vittorio Rampa, and Umberto Spagnolini. Hidden markov models for radio localization in mixed los/nlos conditions. *Signal Processing, IEEE Transactions on*, 55(4):1525–1542, 2007.

[Mpl] Matplotlib.

[MR07] E Mok and Günther Retscher. Location determination using wifi fingerprinting versus wifi trilateration. *Journal of Location Based Services*, 1(2):145–159, 2007.

[NE09] A. S. Pentland N. Eagle. Eigenbehaviors: identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63, 2009.

[Pan] Pandas - python data analysis library.

[Par] Pardus game.

[PBKA08] Andrey Tietbohl Palma, Vania Bogorny, Bart Kuijpers, and Luis Otavio Alvares. A clustering-based approach for discovering interesting places in trajectories. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, SAC '08, pages 863–868. ACM, 2008.

[PSL] The python standard library.

[Rab89] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, Feb 1989.

[RCC+04] Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2658–2663, 2004.

[Ros09] S. M. Ross. *Introduction to probability models.* Academic Press, 2009.

[RS14] M. Szell R. Sinatra. Entropy and the predictability of online life, 2014.

[RSH+08] Injong Rhee, Minsu Shin, Seongik Hong, Kyunghan Lee, and Song Chong. On the levy-walk nature of human mobility. In *INFOCOM 2008. The 27th Conference on Computer Communications. IEEE*, April 2008.

[SCL03] H. C. Lu S. C. Liou. Applied neural network for location prediction and resources reservation scheme in wireless networks. *International Conference on Communication Technology Proceedings*, 2, 2003.

[SL] Scikit-learn.

[SQBB10] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[SSS+14] Arkadiusz Stopczynski, Vedran Sekara, Piotr Sapiezynski, Andrea Cuttone, Mette My Madsen, Jakob Eg Larsen, and Sune Lehmann. Measuring large-scale social networks with high resolution. *PloS one*, 9(4):e95978, 2014.

[Tho62] A Thomasian. Review of 'transmission of information, a statistical theory of communications' (fano, r. m.; 1961). *Information Theory, IRE Transactions on*, 8(1):68–69, January 1962.

[TSA09] C. A. V. Campos L. F. M. de Moraes T. S. Azevedo, R. L. Bezerra. An analysis of human mobility using real traces, 2009.

[UNW] World population to 2300.

[WLP] Ieee 802.11 - wireless lans.

[XL13] N. Bharti A. J. Tatem L. Bengtsson X. Lu, E. Wetter. Approaching the limit of predictability in human mobility, 2013.

[YA05] Moustafa Youssef and Ashok Agrawala. The horus wlan location determination system. In *Proceedings of the 3rd International Conference on Mobile Systems, Applications, and Services*, MobiSys '05, pages 205–218. ACM, 2005.

[YYZS10] Shunsen Yang, Xinyu Yang, Chao Zhang, and Evangelos Spyrou. Using social network theory for modeling human mobility. *Network, IEEE*, 24(5):6–13, 2010.

[ZF12] Nan Zhang and Jianhua Feng. Polaris: A fingerprint-based localization system over wireless networks. In *Web-Age Information Management*, pages 58–70. Springer, 2012.

[ZFL+07] Changqing Zhou, Dan Frankowski, Pamela Ludford, Shashi Shekhar, and Loren Terveen. Discovering personally meaningful places: An interactive clustering approach. *ACM Trans. Inf. Syst.*, 25(3), July 2007.

[ZG10] M. Zignani and S. Gaito. Extracting human mobility patterns from gps-based traces. In *Wireless Days (WD), 2010 IFIP*, pages 1–5, Oct 2010.