Identifying movement patterns from large scale WiFi-based location data

A case study of the TU Delft Campus

Balazs Dukai
Simon Griffioen
Matthijs Bon
Martijn Vermeer
Xander den Duijn
Yuxuan Kang

TU Delft

Identifying movement patterns from large scale WiFi-based location data

A case study of the TU Delft Campus

by

Balazs Dukai
Simon Griffioen
Matthijs Bon
Martijn Vermeer
Xander den Duijn
Yuxuan Kang

Synthesis project of Geomatics at the Delft University of Technology,

Project duration: April 16, 2016 – June 17, 2016

to be presented on Friday May 20, 2016.

TU Delft

Preface

Balazs Dukai
Simon Griffioen
Matthijs Bon
Martijn Vermeer
Xander den Duijn
Yuxuan Kang
Delft, May 2016

Contents

1

Preface

During the fourth quarter of the first year of the MSc Programme Geomatics for the Built Environment at the TU Delft, the Geomatics Synthesis Project (GSP) takes place. This report is part of this framework and in this project, students will apply all their knowledge they have acquired during the courses while working in groups of five or six students. The students will gain experience throughout the entire process of project management, data processing, data analysis, application and presentation.

This year, the GSP focusses on Wi-Fi tracking data from the eduroam network of the TU Delft. The student will be divided into three groups, each researching one of three different topics:

- Identifying occupancy
- Identifying movement patterns
- Identifying activities

This project is dedicated to the second topic, identifying movement patterns. The project requires 3 main documents: ***1***) the baseline review; ***2***) the mid term review, and ***3***) the final review This document embodies the final review and was created to provide the students, the supervisor(s) and other involved parties with an overview of the project. The document includes the problem description, development process, results, conclusions and recommendations for future work.

Delft, University of Technology
June, 2016

2

Glossary

Used terms and abbreviations:

Faculty names

AE / LR	Aerospace Engineering
TNW	Applied Sciences
BK	Faculty of Architecture
CiTG	Civil Engineering
EGM	Thermal Power Plant
EWI / EEMCS	Faculty of Electrical Engineering Mathematics and Computer Science
HSL	Hypersonic Wind Tunnel
ID / IO	Industrial Desgin
FMVG / FMRE	Facility Management & Real Estate
ISD	International School Delft
LMS	Logistics and Environmental Services
LSL	Low Turbulence Tunnel
O&S	Onderwijs & Studentenzaken
RID	Reactor Institute Delft
SC	Sport Center
TPM	Technology, Policy and Management

3

Executive Summary

4

Introduction

4.1. Intro

Wireless Local Area Networks (WLAN) are widely used for indoor positioning of mobile devices within this network. The use of the Wi-Fi network to estimate the location of people is an attractive approach, since Wi-Fi access points (AP) are often available in indoor environments. Furthermore, smart phones are becoming essential in daily life, making it convincing to track mobile devices. This provides a platform to track people by using Wi-Fi monitoring technology. Knowledge of people's locations and related routine activities are important for numerous activities, such as urban planning, emergency rescue and management of buildings.

To understand the human motion behaviour many studies are conducted based on data collection of GPS receivers. The Global Navigation Satellite System (GNSS) is commonly used to track people in large scale environments. However due to poor quality of received signals from satellites in urban or indoor environments, GNSS receivers are not suitable in these environments. Moreover, GNSS receivers are convenient for self-tracking, but for large scale movement analysis, this data should be made available first before others can use it. This led to the development of alternative technologies to track people's locations, including Bluetooth, Dead Reckoning, Radio frequency identification (RFID), ultra-wideband (UWB) and WLAN (Mautz 2012). WLAN has the advantage of widespread deployment, low cost and with the use of a smartphone as a receiver, the possibility to track a large amount of people.

In general, there are four different location tracking techniques by using the Wi-Fi network: Propagation modelling, multilateration, Fingerprinting and Cell of Origin (CoO). Many of these methods rely on Received Signal Strength Indicators (RSSI) and/or previous set of calibration measurements. In comparison, CoO is the most straightforward technique and uses the location of the AP, to locate the mobile device. For, the location of the AP a mobile device is connected to, will give an estimation of the mobile devices' location, and thus the person. For this project, APs related to buildings and building-parts are used to track people's movement.

At the Technical University of Delft (TU Delft) a large scale Wi-Fi network is deployed across all facilities covering the indoor space of the campus. The network is known as an international roaming service for users in educational environments and called the eduroam network. It allows students and staff members from one university to use the infrastructure throughout the campus for free. This allows for large scale collection of Wi-Fi logs including individual scans of mobile devices. A continuous collection of re-locations of devices to access points for a long duration will return detailed records of people's movement. This ubiquitous and individual history location data derived from smartphones will present valuable knowledge on movement on the campus. For this reason, the project is carried out in request of the University's department of Facility Management and Real Estate (FMRE).

In this project, Wi-Fi monitoring technology is used to discover movement patterns on the campus of TU Delft. Based on the relationship between activities and places, location history can be used to discover significant places, movement patterns and hotspots. FMRE can use this information to answer questions such "what is the relation between buildings", "where do people come from" and "how regular a trajectory occurs".

This project will present a method for identification of movement patterns in a large scale indoor environments and between buildings. The method uses concepts of sequential pattern mining. Previous research has been done on sequential pattern mining, such as Zhao et al. 2014 to discover people's life patterns from mobile Wi-Fi scans, Meneses and Moreira 2012 analysed place connectivity using the eduroam network and Radaelli et al. 2013 identifies indoor movement patterns by analysing a sequence of relocations. Individual movement can be identified as a sequence of relocations of a mobile device to different APs. Without any data between two subsequent re-locations, sequential analysis is a convincing way for identifying moving patterns from wifilogs.

4.2. Purpose statement

The project is initiated by the idea that communication technologies can also be used to collect information about connections and connection attempts to Access Points (APs). This geo-referenced information can potentially be used to: ***1***) estimate the number of devices at a location at a certain time, representing presence of people at that place at that time or for a certain duration; ***2***) track unique ID's over several APs, reconstructing individual patterns of movement, resulting in aggregated flows of people and; ***3***) define regular and irregular (temporal, deviating) activities at specific places.

This research will focus on the second matter. Identifying movement patterns has attracted significant interest in recent years. Numerous methods have been explored including Wi-Fi tracking. This report will explain how movement patterns can be identified using large scale Wi-Fi based location data, and tries to contribute with four proposes. ***1***) A method for identifying movement patterns by analysing individual sequences of relocations from a large scale Wi-Fi network; This includes filtering the raw data and automatically create individual trajectories over a time interval as a sequence of relocations; ***2***) Identify spatio-temporal movement patterns of large crowds of people; ***3***) Investigate different visualization methods for showing movement, based on a large scale Wi-Fi network. ***4***) A method for analysing indoor movement using a constructed network graph of the underlying building floorplan.

The contributions can be described in one research question for this project.

- How can movement patterns be identified from large scale Wi-Fi-based location data of the eduroam network?

In order to answer the research question, there are three applied subquestions:

- What patterns can be identified moving from and to the TU Delft campus?
- What movement patterns can be identified between buildings on TU Delft campus?
- What movement patterns can be identified between large indoor regions of the Faculty of Architecture?

Besides looking at this project from a spatial pattern perspective, this research also aims to investigate the following topics:

- Privacy – how viable is the data for personal concerns?
- Validity & Accuracy – how reliable is the data, how accurate, how robust for errors?
- Representativeness – which amount of the actual users is covered? Is this ratio constant or location dependant?
- System of APs – how well is the system equipped for measuring and tracking, and what is missing /essential to use the system this way?

4.3. Methods

The Geomatics Synthesis Project (GSP) is a small research project that combines a literature study with practical research. This includes a case study of the TU Delft campus, using real-world data. Practical work includes data storing, processing, analysing, interpretation, visualization and validation. The project is carried out in a team of six students with a connection to a supervisor and stakeholders (FMRE). This involves interactive discussions between stakeholders as an important part of the research.

4.4. Top level requirements

To keep track of the progress of the project, it is necessary to monitor to which degree the project is meeting the top level requirements and if the project is still on schedule with these requirements. In the baseline review the requirements are specified using the MoSCoW rules and killer requirements. In this chapter these requirements will be discussed and changes will be explained.

MUST building level

- Main goal: Identify movement patterns and connectivity between building entrances.
- Relate entrances (place) of buildings to the corresponding APs (location).
- Find the stay places of each individual in order of the scan time.
- Find individual trajectories from a sequence of stay places.
- Find the movement patterns, by deriving a sequence of common places shared by all trajectories.
- Visualize the movement patterns between buildings in static maps.

A killer requirement for this level is:

- Identification of APs relating to an entrance of a building

SHOULD buildingpart level

- Main goal: Identify movement patterns between large indoor regions.
- Create a network graph from the underlying building floorplan for the analysis, where each region is a node.
- Find the movement trajectories between regions as a sequence of stays.
- Find the movement patterns between large indoor regions.
- Visualize the movement patterns between regions of buildings.

The killer requirements for this level are:

- Digital indoor floorplan of the buildings with classified/named regions (e.g. study rooms, canteen, etc.)
- Georeferenced building floorplans with APs.

COULD room level

- Main goal: Classification of movement patterns at room level.

The killer requirements for this level are:

- Digital indoor floorplan of the buildings with classified/named rooms (offices, classrooms, project studios, corridors, etc)
- Location of access points
- Fingerprinting map

The following chapters will reflect on these requirements, indicating how successful the project is.

4.5. Reading guide

5

Context

5.1. Use case: TU Delft (working title)

This projects main area of interest is the campus of the TU Delft. There are more than 20.000 students using the campus on more than 150 hectares. This emphasises even more the magnitude of this project. The network logs the devices connected to the eduroam access points, which implicitly means logging the (approximate) location of the person carrying the device and more information. This tracking data can be used to derive information about the personality of the person carrying the device, such as the distinction between staff and students, based on the tracked locations. Connection to the Wi-Fi eduroam network is free of charge and requires only a NetID, which all students and staff get upon registration at the university.

It is very important to understand, that 'no data is also data'. This means that a devices that is not being tracked by any access point for a period of time, is either off-campus or disconnected and still on campus. This provides valuable information when researching the movement patterns. This will be further discussed in the chapter 8.

The eduroam network of the TU Delft campus consists of 1730 access points, distributed over more than 30 buildings. The data is collected for each of the access points over a period of little more than 3 months. The logs are stored in a database on a virtual server, where it is accessible to the three project groups and the Geomatics staff. The data that is collected and the storage in the database is further described in subsection 5.6.1.

The department of Facility Management and Real Estate (FMRE) is the main client for the entire Synthesis Project. They would like to know how the campus is being used, what the hotspots on campus and in buildings are, when people travel the most from one building to another and which buildings are most visited.

5.2. Previous research: Rhythm of the campus

In the fall of 2014, similar research was conducted during another edition of the Geomatics Synthesis Project. The group "Rhythm of the campus" investigated the use of the Library and the Aula of the TU Delft, to gain insight in patterns the use of the facilities of the Library and Aula. This section will give a short summary of their research (Van der Ham et al. 2014).

During the project, the group used passive Wi-Fi monitoring to detect users of the TU Delft Library and the Aula to gain insight in the occupation, in request of FMRE. They used BlueMark sensors at the Library, Aula and 5 other faculties for a period of one week and collected ground truth data for 2 days. Due to its sheer size, the raw data was difficult to process. The data was filtered from static devices and outliers and the data analysis resulted in identification of the occupation of the Library and the Aula. The end results was a dashboard which visualized the sensor network, data analysis and pattern recognition to help the client in the decision making process.

This research was different from the research conducted in this Synthesis Project, mainly due the larger size of the eduroam network and the ability to track everybody using the Wi-Fi network.

5.3. Privacy

5.4. Data validity and accuracy

5.5. Representativeness

5.6. Data description and System of APs

5.6.1. Data description

This section will describe the main datasource within the Synthesis Project; a PostgreSQL database containing the logs from the Wi-Fi scanners on the TU Delft campus. Each row in the wifilog table provides a data value for each column (Table 5.1).

username	mac	asstime	apname	maploc	sesdu
j85cCQrGQa..	l6iOu+xxOq..	14-4-2016 12:30	A-23-0-029	System Campus >23-CITG >4e Verdieping	1:32:0
wrBqMLL+j..	f2Pw/PMb4/..	14-4-2016 7:49	A-23-0-035	System Campus >23-CITG >5e & 6e Verdieping	5:32:1
wrBqMLL+j..	f2Pw/PMb4/..	14-4-2016 13:22	A-23-0-035	System Campus >23-CITG >5e & 6e Verdieping	0:40:2
wrBqMLL+j..	f2Pw/PMb4/..	14-4-2016 14:02	A-23-0-093	System Campus >23-CITG >5e & 6e Verdieping	1:27:1
wrBqMLL+j..	f2Pw/PMb4/..	14-4-2016 15:29	A-23-0-091	System Campus >23-CITG >5e & 6e Verdieping	0:05:0
wrBqMLL+j..	f2Pw/PMb4/..	14-4-2016 15:34	A-23-0-035	System Campus >23-CITG >5e & 6e Verdieping	1:42:3
J0IwA+vpcS..	HkLY1UMpBV..	14-4-2016 11:33	A-23-0-035	System Campus >23-CITG >5e & 6e Verdieping	1:27:4
J0IwA+vpcS..	HkLY1UMpBV..	14-4-2016 13:01	A-23-0-035	System Campus >23-CITG >5e & 6e Verdieping	1:01:0
J0IwA+vpcS..	HkLY1UMpBV..	14-4-2016 14:02	A-23-0-035	System Campus >23-CITG >5e & 6e Verdieping	3:30:1
J0IwA+vpcS..	HkLY1UMpBV..	14-4-2016 17:32	A-23-0-035	System Campus >23-CITG >5e & 6e Verdieping	0:40:0

Table 5.1: A segment of the main datasource; the wifilog table

The data value for each attribute (column) in the wifilog table will be described in more detail.

username:
The username column provides the username, as a hashed text. Every user has a unique username, but can appear in the data more than once.

mac:
The mac column provides the media access control adress (MAC address), as a hashed text. The MAC address is a unique identifier assigned to a specific piece of hardware, such as the network adapter located in Wi-Fi devices (mobile phones, tablets, laptops etc.). So, it would be possible that a user can have more than one device connected to the Wi-Fi eduroam network.

asstime:
The asstime is the time of which a connected device is recorded by the system.

apname:
The apname is the name assigned to the access point. Every access point has a unique name.

maploc:
The maploc describes the location of the access point. There could be multiple access points with the same maploc. For instance, there are 31 access points located on the ground floor of the Faculty of Architecture.

sesdur:
The sesdur describes the session duration of which a device is connected to the access point. Because this is not as straightforward as it seems, this will be explained more extensively.

bladibla figure shows the the frequency of session durations.

picture

The figure shows a large peak at exactly 5 minutes, a peak at approximately 5 minutes and decreasing peaks after a time interval of approximately 5 minutes. It looks like it is recording in a certain time interval in which the device is (still) connected.

In order to justify this, the query below is used to see the asstimes (and time to next asstime)

```
select *, asstime_next-asstime as difference
from (
        select count(*),asstime, lead(asstime) over (order by asstime) asstime_next
        from wifilog
        where extract(day from asstime) = 4
        and extract(month from asstime) = 4
        and extract(year from asstime) = 2016
        group by asstime
        order by asstime) as subquery
```

count	asstime	asstime_next	difference
2578	4-4-2016 13:04	4-4-2016 13:09	0:05:10
2435	4-4-2016 13:09	4-4-2016 13:15	0:05:11
2486	4-4-2016 13:15	4-4-2016 13:20	0:05:11
2530	4-4-2016 13:20	4-4-2016 13:25	0:05:11
2471	4-4-2016 13:25	4-4-2016 13:30	0:05:11
2444	4-4-2016 13:30	4-4-2016 13:35	0:05:11
2524	4-4-2016 13:35	4-4-2016 13:40	0:05:11
2588	4-4-2016 13:40	4-4-2016 13:46	0:05:12
2690	4-4-2016 13:46	4-4-2016 13:51	0:05:11
2560	4-4-2016 13:51	4-4-2016 13:56	0:05:11

Table 5.2: The time and time to next scan at a random day

Table 5.2 shows that the time to the next scan is 5 minutes and several seconds in all cases. Most important is to know that all access points are recording the connected device(s) is at the same time.

Table 5.3 will be used to explain the way the time interval of approximately 5 minutes is coming back in the session duration.

id	**username**	**mac**	**asstime**	**apname**	**maploc**
1	oHh0SzXKc0..	WWW0Cdd++v..	1-4-2016 10:13	A-12-0-104	System Campus >12-Kramerslab & Proeffabriek >!e Verd
2	oHh0SzXKc0..	WWW0Cdd++v..	1-4-2016 10:18	A-132-0-064	System Campus >32-OCP-IO >1e Verdieping
3	oHh0SzXKc0..	WWW0Cdd++v..	1-4-2016 11:36	A-132-0-105	Root Area
4	oHh0SzXKc0..	WWW0Cdd++v..	1-4-2016 11:51	A-132-0-066	System Campus >32-OCP-IO >1e Verdieping
5	oHh0SzXKc0..	WWW0Cdd++v..	1-4-2016 14:01	A-132-0-069	System Campus >32-OCP-IO >1e Verdieping
6	oHh0SzXKc0..	WWW0Cdd++v..	1-4-2016 14:06	A-132-0-133	System Campus >32-OCP-IO >4e Verdieping
7	oHh0SzXKc0..	WWW0Cdd++v..	1-4-2016 14:12	A-132-0-066	System Campus >32-OCP-IO >1e Verdieping
8	oHh0SzXKc0..	WWW0Cdd++v..	1-4-2016 14:17	A-132-0-104	System Campus >32-OCP-IO >2e Verdieping
9	oHh0SzXKc0..	WWW0Cdd++v..	1-4-2016 14:22	A-132-0-067	System Campus >32-OCP-IO >1e Verdieping
10	oHh0SzXKc0..	WWW0Cdd++v..	1-4-2016 14:27	A-132-0-066	System Campus >32-OCP-IO >1e Verdieping

Table 5.3: Varying session durations

The first record shows the device is not connected to any of the access points on the campus in the subsequent moment of recording, resulting in a session duration of exactly 5 minutes. In the last record in Ta-

ble 5.3 shows the result of a device that is still connected to the same access point at the subsequent moment of recording. In this case the session duration will be 10 minutes and 21 seconds. This is the time interval between the first moment the device is recorded and the first time the device is not recorded by the same access point anymore. The record with id number 6 describes a situation in which the device is connected to an access point at the moment of recording and connected to another access point at the subsequent moment of recording, the session duration is 5 minutes and 18 seconds in this case. This is the time interval between the two moments of recording.

snr:
The signal to noise (snr) describes a measurement that compares the signal strength to the level of background noise (in dB).

rssi:
The received signal strength indicator (rssi) describes the received signal strength (in dB).

5.6.2. System of APs

This section will describe the current layout of access points (APs) on the TU Delft campus. The exact location of APs in a building is not known, but for the Faculty of Architecture. Therefore the system of APs in the Faculty of Architecure will be described in more detail.
The total number of access points, distributed over more than 30 buildings. Mostly placed on walls or ceilings. Every access point has a floor location, but this does not mean only people on that floor can connect to that access point. Also rooms with high ceilings, such as the orange hall in bk, can have access point located at first floor level but also serve people

6

Movement patterns

6.1. Introduction

The objective of this project is to identify movement patterns. To have a better understanding of this concept, it is important to describe relevant types of movement patterns in a systematic and comprehensive way. A classification of different patterns will provide guidelines for development of different mining algorithms and identify patterns. This chapter will first approach the definition of movement patterns. Subsequently, the theory is demonstrated with the research case of TU Delft. This illustrates what type of pattern mining methods can be used on a movement dataset.

6.2. Movement Identification

By definition, moving objects are entities whose positions of geometric attributes change over time (Dodge, 2008). People always move in geographic space, this means that human movement is geo-referenced. When the start and end time of one movement is specified, its trajectory can be constructed by ordering several movements of one individual. These trajectories can be visualized and analysed.

In order to identify movement patterns, it is important to understand what types of patterns may exist in the data. Besides, there are many types of patterns and not everything is relevant for this project. Therefore, this section will organize various categories. This project aims to identify three different movement patterns: ***1***) Spatio-temporal movement patterns; ***2***) ordered co-location in space; ***3***) unordered co-location in space.

Individual and group movement

Patterns can occur in individual movements or in movements of a larger group. Typical movements of individuals will be different from typical movements of a larger group. For analyzing movement in a larger area with more than 25.000 users, we are interested in typical movement at the larger aggregate level of crowds.

6.2.1. Spatio-temporal movement patterns

As described previous in this section, movement is from one location, or state, to another state, i.e. A to B. These movements can be analysed from movement data to detect the direct connectedness and flow between two locations in a time interval. Questions such as “where do people come from” and “how many people move between two locations” can be answered. Several patterns can be identified from this analysis. Firstly, the number of movements over time can be detected. This will provide insight in the behaviour of humans, e.g. when people go home or at what time people have lunch. Secondly, the flow and direction between two states, i.e. the analysis of the direction of the flows provides information on the symmetry of movement between two locations. For example, if movement 100 people move from A to B within a time interval and 100 people move from B to A in the same time interval, the movement pattern is perfectly symmetrical. Besides analysing movements between two states, consecutive movements of one individual can be used to identify movement patterns. These trajectories will be the basis for the next section to identify co-locations of several trajectories.

6.2.2. Co-location in space

When moving individuals share some locations in their trajectory, you can speak of co-location in space. According to Dodge (2008) there are three types of co-location in space: ***1***) ordered co-location occurs when some locations are shared by multiple trajectories in the same order; ***2***) unordered co-location when shared locations are attained in different orders; ***3***) symmetrical co-location when the shared locations are in opposite order. This means that co-location in space, helps to identify movement patterns in the sense of frequently visited locations in one trajectory. For example buildings A, B, C can be visited in the same order by multiple trajectories, and the same buildings can be visited by multiple trajectories, but in different orders.

Ordered co-location is space can be analysed with the concept of sequences. A sequence is an ordered list of visited locations. Sequential pattern mining algorithm help to understand a what order common locations are visited. In this report, trajectories of a sequence of locations are analysed to identify ordered co-location in space movement patterns. Unordered co-location in space analyses the same trajectories, but does not consider direction or order of the movement. This means that common locations visited together in one trajectory can be identified. In other words, the association between buildings is detected. A commonly used method to detect groups of objects in a list (i.e. a trajectory), an association rule mining algorithm is used. This report will use the concept of this algorithm to identify these groups of buildings that are frequently visited together.

7

Methodology

8

Preprocessing

Before movement patterns between buildings can be retrieved, pre-processing of the raw data is required. In this chapter the different pre-processing steps will be described in detail. First section 8.1 addresses the initial data filtering. section 8.2 describes the filling of the dataset with a 'world' location, this enables detection of movement from and to the campus..section 8.4 is about filtering of records of people only passing by a building. Finally section 8.3 concerns the grouping of records of the same mac address that are subsequent in time and at the same location.

8.1. General filtering

Each record in the wifilog represents the scanning of a certain device at a certain time by a certain access point. In order to detect the movement patterns of these devices between buildings it should be known for each access point in which building it is located. The apname field in the wifilog table includes the building id in which building each scanner is located. However for some access points the apname is given in a different format and as a result their location is unknown. These apnames have in common that they don't contain the '-' character which is present in all the other apnames. As a result the apnames of which the location is not known can simply be filtered out by checking if a '-' is present in the apname.

8.2. Filling

Because the dataset contains all records of when certain devices are scanned, it also Implicitly stores information on when the device is not located at the campus. These time gaps in which a particular device is not scanned at the campus give information on when the corresponding person is not at the campus. This information is valuable for detecting movement patterns from and to the campus in addition to the movement between buildings at the campus. Considering the fact that many student only visit one faculty each day. It becomes especially clear, that the movement from and to the campus plays an important role in the overall movement pattern of a person. In order to be able to directly derive movement from and to the campus from the dataset, the time gaps present in the data should be stored explicitly. Therefore each time gap larger than an hour is filled with a 'world' record. The word 'world' is used to indicate that the device could be located at any place in the world during the time spans that it is not scanned at the campus. The begin and end time of a world record is defined by the end of the previous record and the start of the next record in time. In case there is no previous or next record the boundaries are defined by the starting time of the whole dataset and the current time. Figure 8.1 visualizes the filling of time gaps for one devices. The black intervals indicate the time during which a device was scanned at the campus, the red intervals indicate the time gaps filled with a world records. Note that the gap at 16:00 is smaller than an hour and therefore is not filled.

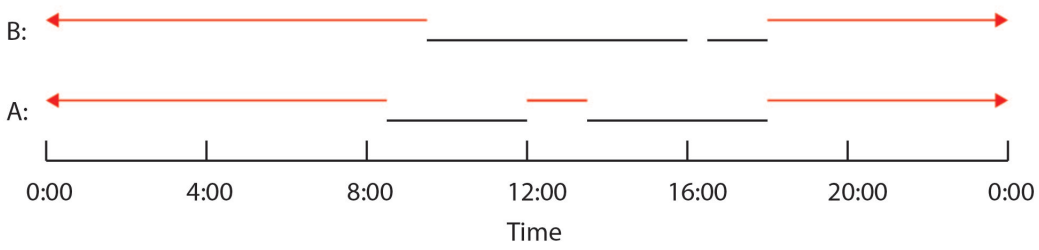

Figure 8.1: Filling

8.3. Grouping

In order to reduce the data and to be able to filter on people only passing by a building without going in, the data needs to be grouped. The goal is to identify movement patterns between different buildings, this means that records of subsequent scans of the same device in the same building can be grouped together into one single record. The mobile of someone who for example studies the whole day at architecture might have 20 records in the database for that day. This can be reduced to one record that states the time the device arrived at Architecture and left again. To determine whether two records are subsequent in time, and therefore should be grouped together, a threshold for the time gap between two records needs to be defined. As explained in section ... the eduroam system has 'scanning rounds' at intervals of 5 minutes and several seconds. If a device is not scanned during a scannig round, but was scanned the round before, the end time of the records is set to the time of the previous scan round plus 5 minutes (see record A1 and B1 Figure 8.2). As a result the gap will be a bit more than 10 minutes if someone is not scanned for 2 subsequent rounds (Figure 8.2 A), and a bit more than 15 minutes if someone is not scanned for 3 subsequent rounds (Figure 8.2 B). It was decided to set the gap threshold or grouping to 15 minutes. The reasoning behind this is that someone who is not scanned for 3 subsequent rounds has likely left the building. For the example this means records A1 and A2 would be grouped together, records B1 and B2 on the other hand are not grouped.

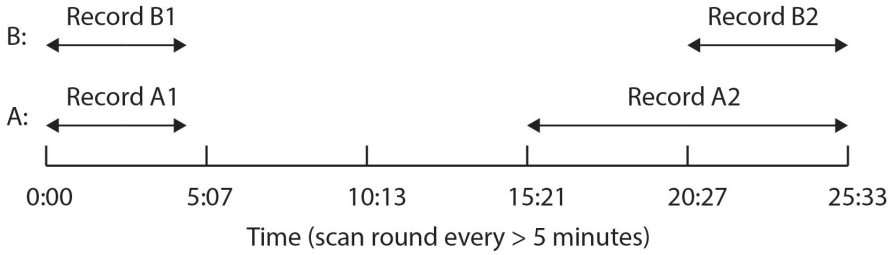

Figure 8.2: Grouping

8.4. Filtering

For the detection of movement patterns between buildings, records of people that only pass by a building without actually visiting it should be excluded. The reason for this is that records of people only passing by a building could result in misinterpretation of the movement patterns. If faculty B is for example located on the route from faculty A to the lunch facility. Then it is likely that people moving from faculty A to the lunch facility are picked up by a scanner located at faculty B. As a result the movement from faculty A to the lunch facility will be visualized via faculty B (see Figure 8.3 top). Someone that isn't aware of the 'passing by' problem might conclude that people from faculty B make most use of the lunch facility. In reality however, people from faculty A make more use of the lunch facility. By filtering out the records of people only passing by buildings the correct movement can be visualized (see Figure 8.3 bottom). It should be noted that filtering out 'passing by' records can only be done after the grouping process. The reason for this is that 5-minute records that would individually be classified as someone passing by might be grouped together. After grouping the combined record is not classified as someone who passes by. Furthermore it should be noted that the filtering of 'passing by' records occurs after filling the data with 'world' records. The reason for this that a passing by event does mean that the device was located on the campus. The world records are meant to represent the

time the device is not on the campus.

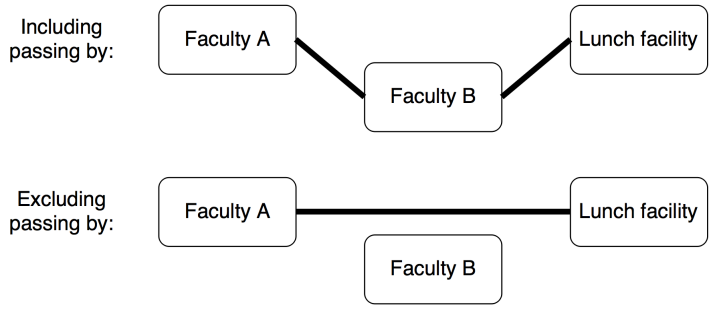

Figure 8.3: Passing by

8.5. Implementation

The filling, grouping and filtering (passing by) steps described above are implemented in an integrated way. The Pseudocode for the implementation is shown in Figure 8.4. As can be seen in the code there is communication with the database at several points. The table from which the records are retrieved for each mac address is already processed as described in the general filtering section. Furthermore the format of the table is slightly different compared to the initial wifilog. The session duration is exchanged for an end time column which is derived by adding the session duration to the asstime (start time of a record).

```
macs = get distinct macs from db
create new empty table with 4 columns(mac, building, start, end)
min_time = minimum time in entire db
max_time = current time
for mac in macs:
        records = get all records for mac from db
        cur_rec = first record from records
        insert world at start (mac, world, min_time, cur_rec[start])                    # fill
        for next_rec in records[1:-1]:
                gap = next_rec[start] - cur_rec[end]
                if gap > hour:
                        insert world (mac,'world',cur_rec[end],next_rec[start])          # fill
                if gap < 15 minutes and cur_rec[building] == next_rec[building]:
                        cur_rec = (mac,cur_rec[building],cur_rec[start],next_rec[end])   # group
                elif cur_rec[end]-cur_rec[start] > 6 minutes:                            # filter passing by
                        insert cur_rec
                cur_rec = next_rec
        if cur_rec[i_end]-cur_rec[i_start] > 6 minutes:                                  # filter passing by
                insert cur_rec
        insert world at end (mac, world, cur_rec[end],max_time)                          # fill
```

Figure 8.4: Pseudocode preprocessing

Figure 8.5 shows an example of the records of one device over a time span of one day during the different pre-processing steps. From the raw data it can be seen that this person spends most of the day in building B. The person is scanned once at building A before he arrives in the morning and after what is likely to be his lunch break. The last two hours the person is scanned in building C. After filling three world records are added, at the beginning of the day, during the lunch break, and at the end of the day. The grouped records show that the subsequent scans in building B and C are grouped together. Finally the scans at building A are removed from the dataset as they are likely to indicate passing by events.

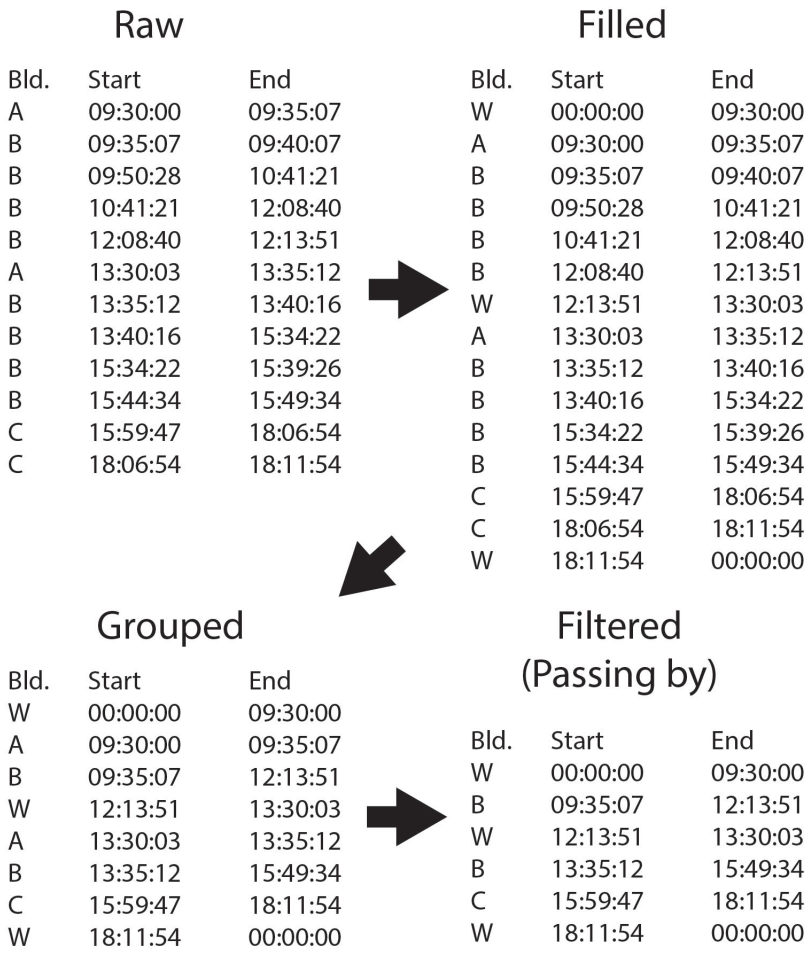

Figure 8.5: Preprocessing

8.6. Apname vs maploc

The data in the table 'wifilog' contains information about the location of the Access Point (AP) in two columns. The first one is the column 'apname', which is a string with the symbolic name of the AP, for example 'A-08-G-010'. The two numbers in the second part of the string, in this case '08', represent the building number. This building number can be linked to a location in the world. The second column which contains information about location, is the column 'maploc'. This column also contains strings, which look as follows:

System Campus > [buildingid] > [specific location]'. An example of such a string is 'System Campus > 21-BTUD > 1e verdieping'. In such a string, the middle part can be linked to a building, so to a real-world location. But there are some other values for maploc, which can less clearly be linked to a real-world location. Such a value is 'Root Area', it is unclear what this value means and it contains no information about a building or area it might be in. This makes it impossible to link it to a location in the world. Then there is the value 'Unknown', a value that indicates that there was no name attached to the Access Point that user was connected to. Again in this case, it is impossible to link this value to a real-world location.

As both 'Root Area' and 'Unknown' are in the minority of records, they could be left out of the queries. But for some records, the column 'apname' did provide information about the location, while the 'maploc' column value was 'Root Area'. In most of these cases however, the building number, the second part of the string, was a number of length three. But there are no buildings on the TU Delft campus with a building number that high. When consulting Wilko Quack about this, he explained that these building numbers had an arbitrary 1 in front of the building number. So 'A-134-A-001' was not building 134, but building 34, which was an actual building number on the campus. This would mean that using the column 'apname' for getting the building number would mean a higher number of results and therefore a more realistic visualization of the movements.

Taking the substring of that column and linking it to a building with an actual location is done in two steps. First the whole string is retrieved and with a function in Python the substring is derived. Subsequently, the building id that is the result of this function can be linked to a table in the database which has for every building five columns: buildingid, name, point (as geometry), x (longitude), y (latitude) (see in **??**).

8.7. Static and mobile devices

In order to identify the movement patterns and know what entrances and exits are most frequently used even better, we aim to identify dynamic and static devices. In our first approach, we will look at the number of different access points the device is scanned by in time. The distinction between static and dynamic devices is important, because the behaviour, in terms of Wi-Fi tracking, is significantly different. For instance, a static device, such as a laptop, connects with the Wi-Fi network at different moments compared to a dynamic device, such as a mobile phone. The difference will be explained more in detail using the image below.

Assume a person that carries a static device (laptop) and a dynamic device (mobile phone) enters a building. While being on his way to the destination, the person does not make use of the laptop, thus the laptop is not connected to the Wi-Fi network. On the other hand, the Wi-Fi of the mobile phone is turned on all the time, and connects at the moment the device is on range of the first access point. On the way the mobile phone is scanned by Access Point(AP) 1, 2 and 3. The person connects to the Wi-Fi network with the laptop at the moment it arrives in the room, of which the Wi-Fi is covered by AP 3. This access point scans the laptop for first time after entering the building. The static laptop is distorting the result, due the fact that in this case the entrance access point for the laptop would be AP 3. In order to achieve a more reliable result, the aim is to filter out the static devices.

To identify the static and dynamic devices, we analyze the behaviour of each device. The first approach focuses on the number of (distinct) access points and the session duration. We assume to find differences between them (Table 8.1).

	Session duration	Nr.of access points
Static	long	low
Dynamic	short	high

Table 8.1: Difference between static and dynamic devices

We expect that the relation between the distinct access points and the (summed) session duration, called ratio, is going to be useful in making the distinction between static and dynamic devices (Equation 8.1).

$$Ratio = distinctaccesspoint/summedsessionduration \tag{8.1}$$

In this, a small ratio indicates the device is dynamic and a large ratio indicates the device is static. The result shows that the number of devices decreases over ratio(Figure 8.6).

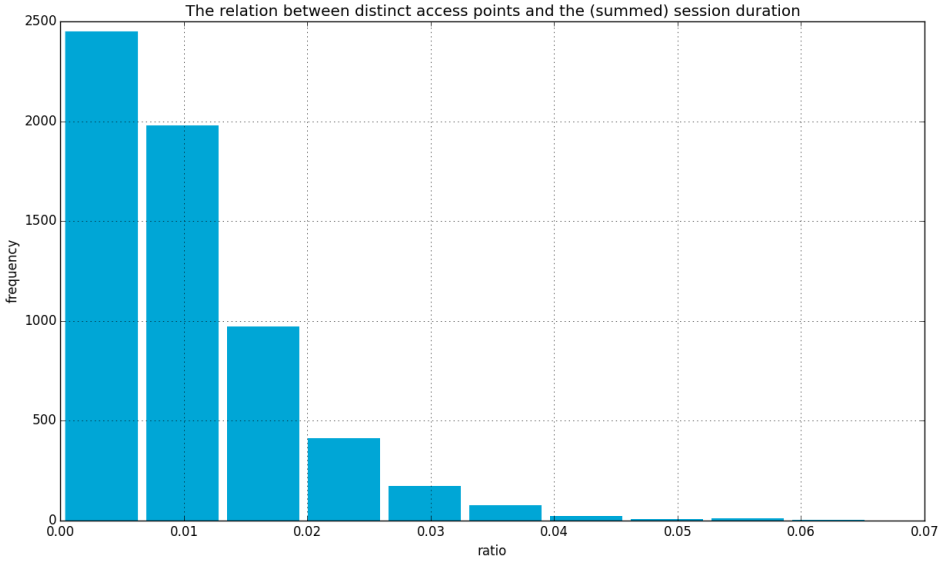

Figure 8.6: The relation indicating frequency of a radio

Because the frequency decreases gradually, there is a fuzzy boundary that separates the static from dynamic devices. Therefore it not (yet) possible to filter out the static devices for further analysis. In order to improve this, the plan is to use the exact number of access points that scanned the device instead of the distinct access points. Also, a closer look will be taken at the session duration, since dynamic devices will have session duration of approximately 5 minutes much more often.

9

Spatio-temperal movement patterns

9.1. Introduction
9.2. Methods
9.3. Results
9.3.1. All movement
9.3.2. Mobile vs static
9.3.3. Week vs weekend
9.3.4. From and to
Campus Library and BK Aula during lunchtime Cantine and bouwpub

10

Trajectory patterns

10.1. Introduction

This GSP attempts to identify people's movement patterns from anonymized wifi logs. chapter 9 described movement patterns including spatial and temporal aspects of single movements of a crowd of people. Another way of looking at movements, is by tracking individual movement for a longer time interval. A large set of individual trajectories can be used for the identification of typical movements among users of the campus. The method uses concepts from sequential pattern mining.

This chapter presents a method for identifying movement patterns using individual trajectories. As described in chapter 6, if moving individuals share some locations in their trajectory, you can speak of co-location in space. When the order of the shared locations are similar for multiple trajectories, you can speak of typical movement. This concept is explored for the identification of movement patterns, and thus the usage of the campus. This approach can answer different questions than looking at single movements, as is done in chapter 9. For example, 'how many places the user frequently visits', 'at what order the user visits places', 'how often a trajectory happens', 'how many places contained in a frequent trajectory'.

First, this chapter will describe the problem description, including the extraction of locations of a user, the mining of individual trajectories from an anonymized Wi-Fi scan list, and finaly the mining of movement patterns from a set of trajectories using the PrefixSpan algorithm.

10.2. Problem description

10.2.1. Location extraction

The data provided by the eduroam network enables a detailed view of people's movement on campus. The large coverage of the eduroam network allows to track users for a large part of the day when they enter the campus. However, the observation space is limited to the extent of the size of the campus, making it not possible to track people outside the eduroam network. A second disadvantage is the spatial resolution of the positioning method. The range a mobile device can be connected to an AP, influences the accuracy of the estimated location of a mobile device. For indoor environments of the TU Delft campus, this is just a few tens of meters wide. This resolution allows tracking movement at a building level by re-locating mobile devices to the closest AP. Data between two re-locations is not available. Therefore, an individual's trajectory is depicted by connecting the re-locations as a sequence of APs. These individual trajectories are used to identify patterns.

A location represents a geographic position where a user stays, i.e. a user is in state. For identifying movement patterns from Wi-Fi monitoring, we are interested in movement between two locations where an individual stays for a longer time period. Such a location, or stay place, can be detected when a user is connected to the same AP for a longer time. To detect buildings as a location (i.e. contains multiple APs), two consecutive Wi-Fi scans must contain APs of the same building. With a data collecting interval of 5 minutes, it means that people will be filtered out if their stay duration is less than 10 minutes. Based on this assumption, people with a shorter stay duration are considered passing by, as explained in chapter 8.

10.2.2. Individual trajectory

An individual's trajectory is constructed as a sequence of locations in order of the scan time. Start and end time of a trajectory can be specified with a time interval, e.g. a day or week. If p is a location, then a trajectory can be written as:

$$p_1 \rightarrow p_2 \rightarrow p_3 \rightarrow \ldots \rightarrow p_n$$

Given a time interval, there is a set of individual trajectories

$$S = \{t_1, t_2, t_3, ..., t_n\}$$

where each

$$t_i$$

is the trajectory over a time interval of one user.

10.2.3. Trajectory Pattern

From a set S of trajectories, different patterns can be identified using sequential pattern mining algorithms. Frequency of a trajectory by all users of the campus can be detected. This can be represented as a trajectory T with a support s. Support means how many times the same sequence, or sub-sequence, is shared in the set of trajectories. This gives valuable information on the order common buildings are used and what order of buildings occurs the most. Furthermore, the length of common trajectories can be discovered. This allows for identification of movement patterns of a specific length n. Also, when location is not considered, but only the length of a trajectory, the mobility pattern of an individual can be described in terms of how many times he/she re-locates.

There exists many developed sequential pattern mining algorithms. For this study PrefixSpan **pei2004mining** is used to identify common shared trajectories or sub-trajectories.

10.3. Implementation

10.4. Results

11

Indoor spatio-temporal movement patterns

11.1. Introduction

As described in the firs part of this report, Wi-Fi tracking data can be used to identify movement between buildings. Given that indoor areas are usually better covered with Wi-Fi access points than outdoor areas, it is natural to also look at movement inside buildings. The following section describes our method of identifying and visualizing indoor movement in the Faculty of Architecture of TU Delft.

The process of indoor movement analysis is conducted along the steps below, thus the section also follows this structure:

1. Delineate building parts based on the layout of access points and the division of the building (e.g. department, canteen, building wing), and group the access point into building parts.

2. Identify movements in the data between building parts.

3. Create a route network that connects the building parts and is constrained on the corridors of the building.

4. Assign the movements to the route network.

5. Visualize the movement along the indoor network.

11.2. Theory / methods

After identifying movement between different buildings, the next level is to do so between different parts inside a building. These parts represent functional or spatial divisions inside a building, e.g. departments, community areas, building wings and are referred to as *building part*.

A prerequisite of the method is to know the at least room level location of the access points in the respective building. At the time when the project was carried out, the detailed access point locations were availably only for the Faculty of Architecture. Thus the focus on this particular building.

As opposed to outdoor pedestrian movement which is not necessarily constrained on a fixed network, indoor movement is constrained by the layout of the respective building. The building parts of the Faculty of Architecture can be represented by its underlying graph, having the building parts as nodes and the corridors as edges Figure 11.1. Then indoor movement is necessarily constrained on this underlying graph.

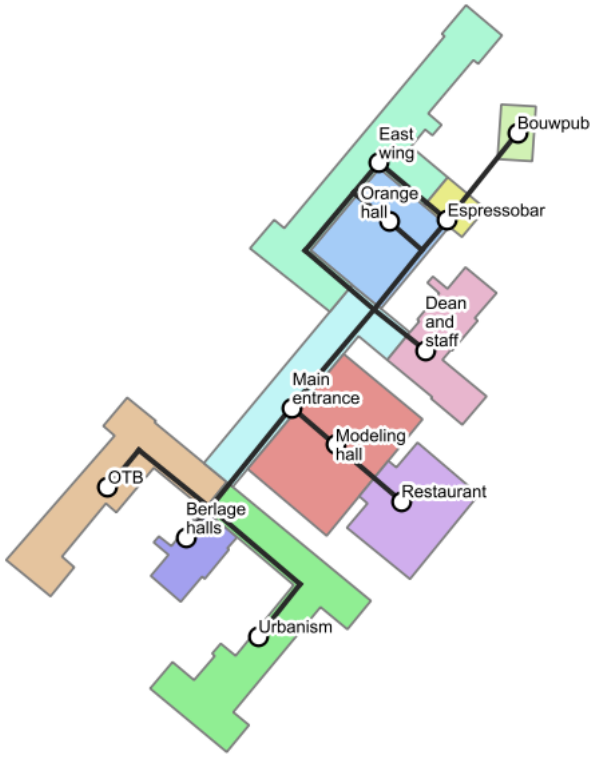

Figure 11.1: Building parts on the ground floor of the Faculty of Architecture and its underlying graph.

The Wi-Fi system of the TU Delft campus has a five minute scan interval, which is too coarse to catch detailed movement indoor. As five minutes is sufficient to reach any two locations in the building taking any route. Therefore not the movement trajectory itself is identified from the data, but the fact of relocation from origin to destination. Then the path of the movement can also be identified by analysing the layout of the building. For example if a person stayed at the Restaurant, then soon after he stayed at the Orange hall, he necessarily had to traverse the corridors in-between these two locations. Our method is based on this assumption.

Due to the building layout, in most of the cases there is only one possible direct route between two building parts. However, in case of multiple route options, the exact route of a movement is assumed to be the shortest route between origin and destination.

11.3. Implementation

Balázs: describing how the map visualization works and implemented, writing in a separate document

11.4. Results

12

Conclusions

First of all it can be concluded from the preliminary results that the Wi-Fi network data is suitable, at least to some extent, for retrieving movement patterns of people. Expected patterns such as a movement peak between building during lunch time, and a morning and afternoon peak of people entering and leaving the campus can be clearly distinguished in the data. Similarly aggregated movement on the map shows the expected result that Aula-Library is the most frequently travelled path. More specific patterns between particular buildings and/or during certain time intervals can easily be derived due to the automated workflow. An example of such a specific pattern is that people moving to the aula most often origin from the faculty of Applied Sciences. Furthermore it can be concluded that Aerospace Engineering and to some extent Architecture are rather isolated compared to the other faculties on the campus. Especially when interpreting the result of movement from and to the campus, it should be taken into account that static devices (mainly laptops) are not filtered yet. Disconnecting a laptop for over an hour will currently still be interpreted as a movement away from the campus and back.

13

Recommendations

13.1. Entrances
13.2. Association rules
13.3. Distinguishing user groups
13.4. Occupancy
13.5. AP system

The setup of the system that logs the devices connected to access points is directly connected to the accuracy of the processed data. Currently, the APs register every device that is connected to it and the logging system receives all connected devices approximately every five minutes. Additionally, all access points are located indoors, logging every device carried by people using that building. These two aspects of the AP system limit the accuracy of the processed data and thus the movement patterns that can be derived.

Because the system logs every connected devices once every five minutes, a device will only be registered if the devices is connected for at least five minutes. This will result in discrepancies in the processed data. Devices and thus people walking by an AP will probably not be registered, for they are not connected to that AP for at least five minutes. This is unfortunate, because a person can travel a rather long distance in five minutes, e.g. making it hard to track people indoors. If the system would be logging every device all the time, irrespective of the time the device is connected, the tracking data would contain every AP that a device would connect to and thus provide much more accurate tracking data. Understandably, logging every user every second would result in huge amounts of data, which would most definitely result in performance issues.

Secondly, because all scanners are located inside buildings, there is little to no information on people when they move from one building to another. Surely something can be told from the time it takes a device from the last scan in one building, to the first scan in the second building. But for outdoor tracking purposes, this system is limited. From some experiments that were conducted on the TU Delft Campus it can be concluded that a device located outdoors near a building can be detected by APs inside the building, but this depends on the antenna in the devices and the exact location of the device in respect to the AP. If more detailed information about movement outdoors is desired, it would be wise to also include outdoor APs in the system.

To improve further research, it is recommended to take the system of APs into account before actually conducting the research. If outdoor movement tracking is desired, outdoor APs are required. And if tracking indoors is one of the goals, the frequency of logging should be set to an interval that is in the order of magnitude of 10 seconds to one minute, taking data size in consideration.

13.6. Data reasoning
13.7. Visual exploration

14

Acknowledgements

We would like to take the opportunity to express our gratitude and regards to everyone that contributed to this project.

First, we would like to thank Edward Verbree, our supervisor, for the feedback provided during the course of the project. Additionally, we would like to thank Wilko Quack for his patience and support with all the issues we encountered when using the database.

We also would like to thank Bart Valks and Iljoesja Berdowski from the department of Facility Management and Real Estate for their feedback and guidance throughout the project and the challenges they proposed.

Additionally we would like to thank Jorge Gill for his feedback on our visualizations and his help for creating even better visualizations.

15

Appendix A

Bibliography

Agrawal, Rakesh, Tomasz Imieliński, and Arun Swami (1993). "Mining association rules between sets of items in large databases". In: *ACM SIGMOD Record*. Vol. 22. ACM, pp. 207–216. (Visited on 04/09/2015).

Agrawal, Rakesh and Ramakrishnan Srikant (1994). "Fast algorithms for mining association rules". In: *Proc. 20th int. conf. very large data bases, VLDB*. Vol. 1215, pp. 487–499. URL: https://www.it.uu.se/edu/course/homepage/infoutv/ht08/vldb94_rj.pdf (visited on 04/09/2015).

Anbukkarasy, G. and N. Sairam (2013). "Interesting Metrics Based Adaptive Prediction Technique for Knowledge Discovery". In: *International Journal of Engineering and Technology* 5.3, pp. 2069–2076. (Visited on 05/16/2016).

Dodge, Somayeh, Robert Weibel, and Anna-Katharina Lautenschütz (2008). "Towards a taxonomy of movement patterns". In: *Information visualization* 7.3-4, pp. 240–252.

Hunter, J. D. (2007). "Matplotlib: A 2D graphics environment". In: *Computing In Science & Engineering* 9.3, pp. 90–95.

Mautz, Rainer (2012). "Indoor positioning technologies". PhD thesis. Habilitationsschrift ETH Zürich, 2012.

Meneses, Filipe and Alberto Moreira (2012). "Large scale movement analysis from WiFi based location data". In: *Indoor Positioning and Indoor Navigation (IPIN), 2012 International Conference on*. IEEE, pp. 1–9.

Radaelli, Laura et al. (2013). "Identifying Typical Movements among Indoor Objects–Concepts and Empirical Study". In: *Mobile Data Management (MDM), 2013 IEEE 14th International Conference on*. Vol. 1. IEEE, pp. 197–206.

Van der Ham, M et al. (2014). "Rhythm of the campus". In:

Zhang, Yuejin et al. (2009). "A Survey of Interestingness Measures for Association Rules". In: *International Conference on Business Intelligence and Financial Engineering, 2009. BIFE '09*. International Conference on Business Intelligence and Financial Engineering, 2009. BIFE '09. IEEE, pp. 460–463. DOI: 10.1109/BIFE.2009.110.

Zhao, Shao et al. (2014). "Discovering People's Life Patterns from Anonymized WiFi Scanlists". In: *Ubiquitous Intelligence and Computing, 2014 IEEE 11th Intl Conf on and IEEE 11th Intl Conf on and Autonomic and Trusted Computing, and IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UTC-ATC-ScalCom)*. IEEE, pp. 276–283.