Identifying movement patterns from large scale WiFi-based location data

-A case study of 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 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	BK City
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. 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 snaps the location of the mobile device to the same coordinate position as the access point it is connected to. For this project, CoO is 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 easy 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, [**?**] 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

Identifying movement patterns has attracted significant interest in recent years. This report will explain how movement patterns can be identified using large scale Wi-Fi based location data. This report tries to contribute with three 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***) Restructure the association rule mining algorithm to use it in a large scale tracking environment, to discover locations that are commonly associated; ***3***) Investigate different visualization methods for showing movement, based on a large scale Wi-Fi network. ***4***) A method for automatically detect what entrances are used to enter and exit a building.

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 well, there are some sub questions:

- What movement patterns can be identified between buildings on TU Delft campus?
- What movement patterns can be identified between large indoor regions on TU Delft campus?
- What entrances are used to enter and exit a building on TU Delft campus?

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 previous requirements will be discussed and possible changes will be explained.

The goals that *must* be achieved are on the level of detail of the campus. It's detailed specification, as stated in the baseline review, is shown below.

MUST campus level Main goal:

1. Identify which entrances are used to enter and exit a building;
2. Identify movement patterns and connectivity between building entrances by sequential pattern mining.

- 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 by taking a time interval 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

Currently the project is progressed so far that it is possible to identify building patterns between buildings. The stay places of individuals and their trajectories have been found and this has been visualized in both static and dynamic maps and bar charts. But, until now there is no accurate map with the location of

all access points of the campus. There is such a map for the faculty of architecture, but it is only one building and not very clear. Until this map of the whole campus becomes available, identifying entrances will be hard to do. section 12.1 will go into greater detail about the progress that has been made so far with entrances.

The goals that *should* be achieved, focus on the building level, where buildings are divided into regions, but since there is currently no map with the locations of the access points, this level of detail is not yet reached. However, the way that the code is setup allows for easy transformation to higher levels of detail when such a map becomes available. How this code exactly works is explained in **??** and **??**.

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 7.

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 section 5.6.

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

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

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 7.1 addresses the initial data filtering. section 7.2 describes the filling of the dataset with a 'world' location, this enables detection of movement from and to the campus..section 7.4 is about filtering of records of people only passing by a building. Finally section 7.3 concerns the grouping of records of the same mac address that are subsequent in time and at the same location.

7.1. 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.

7.2. 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 7.1 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.

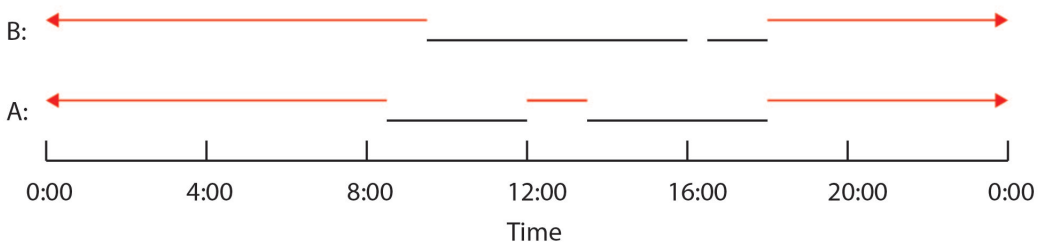


Figure 7.1: Filling

7.3. 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 7.2). As a result the gap will be a bit more than 10 minutes if someone is not scanned for 2 subsequent rounds (Figure 7.2 A), and a bit more than 15 minutes if someone is not scanned for 3 subsequent rounds (Figure 7.2 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.

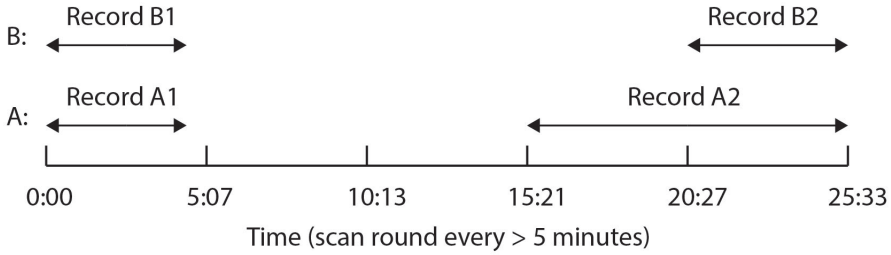


Figure 7.2: Grouping

7.4. 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 7.3 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 7.3 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.

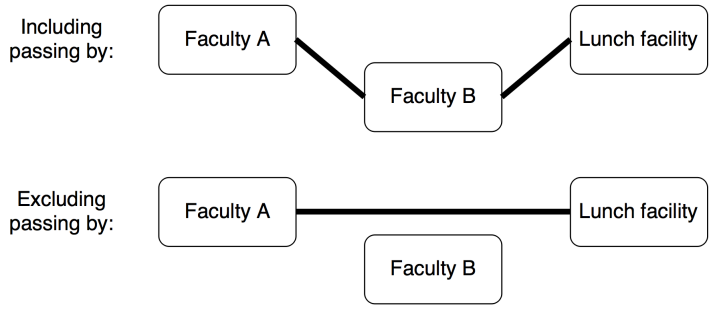


Figure 7.3: Passing by

7.5. 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 7.4. 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 7.4: Pseudocode preprocessing

Figure 7.5 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.

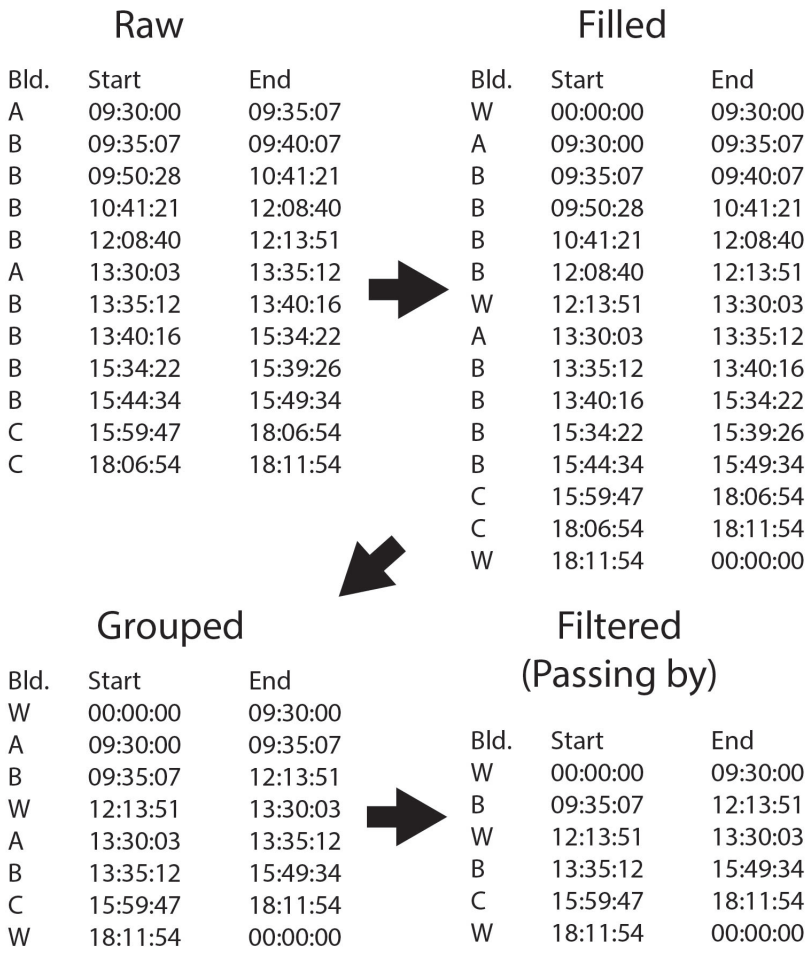


Figure 7.5: Preprocessing

7.6. 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 subsection 8.0.4).

7.7. 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 7.1).

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

Table 7.1: 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 7.1).

$$Ratio = distinctaccesspoint/summedsessionduration \tag{7.1}$$

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 7.6).

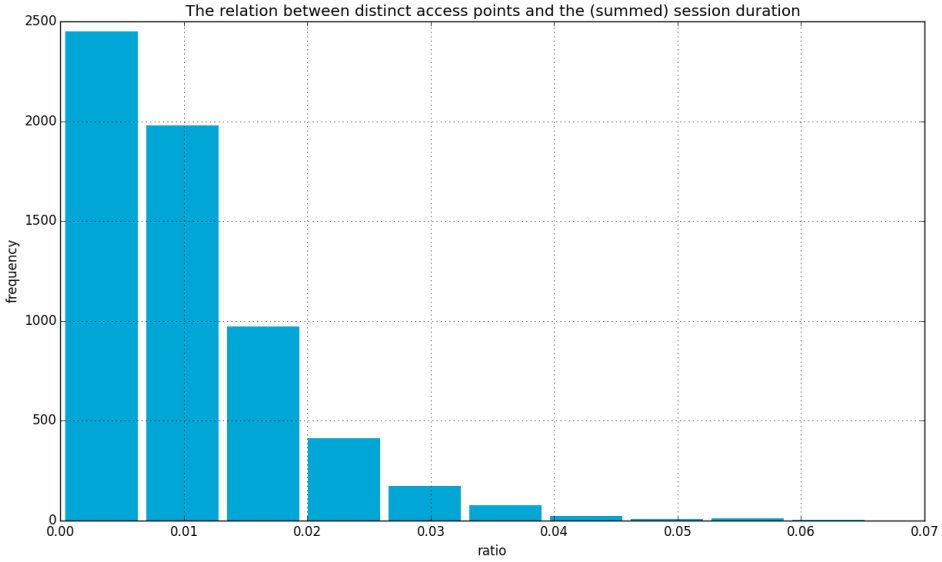


Figure 7.6: The relation indicating frequency of a radio

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.

8

Spatio-temperal movement patterns

To automate the workflow of creating movement visualizations between buildings, a program is created. There is a distinction between two types of visualization:

1. Maps

2. Bar charts

The bar plot visualizes the movement throughout the day in 24 bars. Each bar represents the movement from a selection of buildings to another selection of buildings, over a time interval of one hour. In **??**, the bar charts will be discussed in more detail. For map visualization the JavaScript Leaflet.js is used, this allows for creation of an interactive user interface with a base map from Open Street Maps and visualize the buildings and movement between them. In subsection 8.0.4 the map visualization will be discussed in more detail. For the bar charts the Python module matplotlib was used.

8.0.1. Create movement records

The data resulting from pre-processing contains the states of where a particular device was located during a certain time period. Implicitly this also includes information on the movement of the device. If a device is first located in building A and subsequently in building B it must have moved from building A to B. However, in order to be able to retrieve the movement patterns of devices the movement should be stored explicitly. This means that each record should store the movement of one device from one building to another building or to world. Examples of movement patterns that can be retrieved from this data are: the number of devices moving from building A to B within a given time period, and the peak in movement from building A to all other buildings.

To create records for each movement first the preprocessed data is ordered on mac address and start time. By doing this all the subsequent states for every device are listed directly below each other (see Figure 8.1). As a movement is defined by the change of one state to another, movements records can be created from every two consecutive state records (see Figure 8.1. However, not every two consecutive states represent a movement. Only when the two states concern the same device and they are at different buildings they represent a movement. This means that movement records with different mac addresses or similar building id's are filtered out (see Figure 8.1.

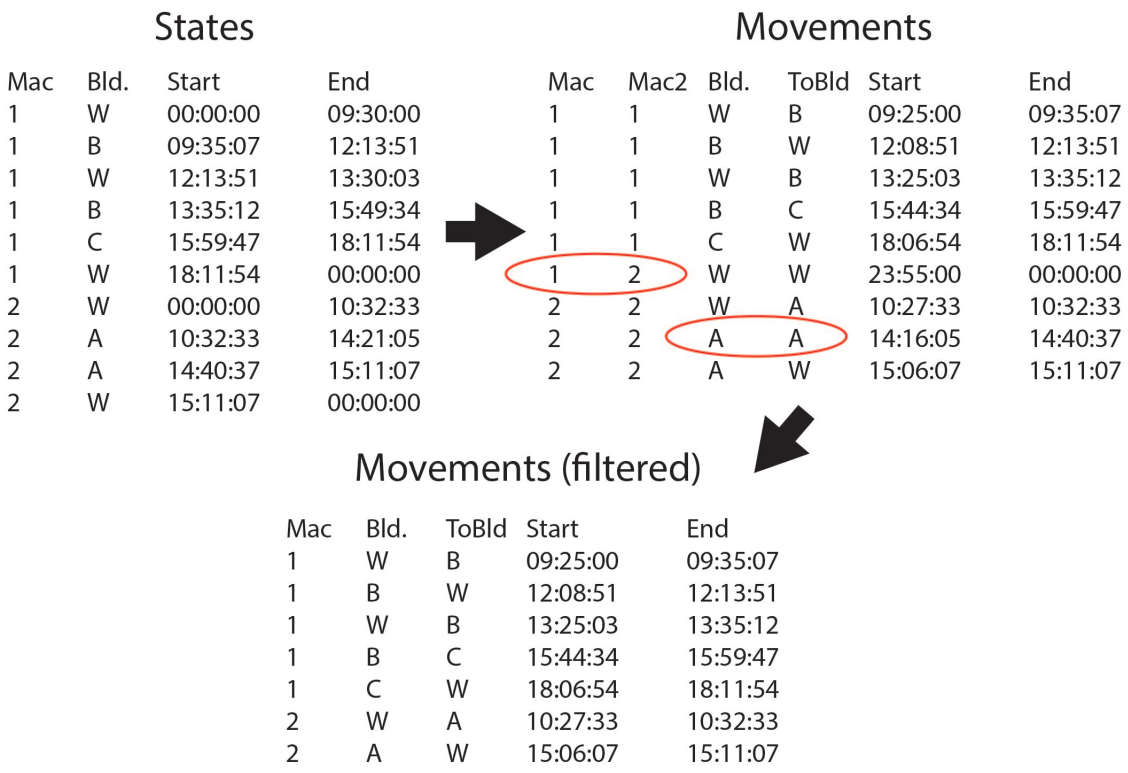


Figure 8.1: Movement records

The start and end time of the movement are defined by the end time of the previous state minus 5 minutes, and the start time of the next state (see **??**). The reason that 5 minutes are subtracted from the end time of the previous state is that this is approximately the last moment in time the device was actually scanned at the location of the previous state. In the figure below the device is scanned 15:21 at building B. Approximately 5 minutes later (at 20:27) the device is scanned at building C. The state record of building B however continues all the way until 20:27, whilst the last time it was actually scanned at building B was 15:21. As a result it can be concluded that the movement from building B to C took place somewhere between 15:21 and 20:27. Therefore the start time of the movement between B and C can be approximated by subtracting 5 minutes from the end time of the state record at B. As can be observed in the movement from A to B is retrieved in the same way.

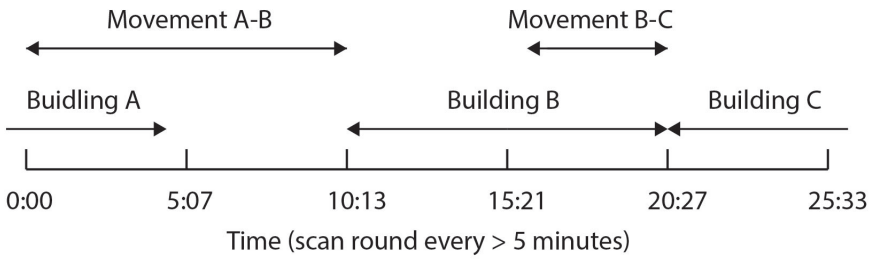


Figure 8.2: Movement

8.0.2. Movement over time

As the described in section subsection 8.0.3 the GUI allows a user to select particular days and specify origin and destination buildings of the movement. Based on this input the movement table can be filtered. Finally the filtered data can be visualized as the amount of movement between the specified buildings at each hour of the day. If the user has specified multiple days, the average amount of movement of these days is taken.

It should be noted that the amount of movement is defined by the number of devices moving between the specified buildings. Figure 8.3 gives an example of the visualization of movement over time.

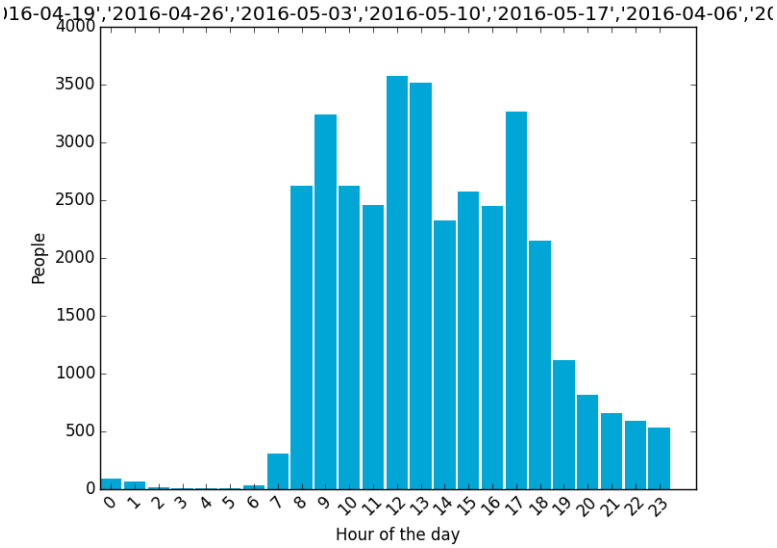


Figure 8.3: Movement overtime

8.0.3. Graphical User Interface

The Graphical User Interface (hereinafter referred to as GUI) for this work is a Python program that shows a Tkinter interface. When the user runs the program, it will display a main window, which is shown in Figure 8.4

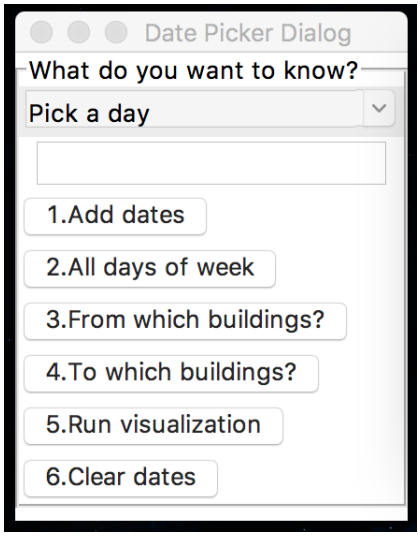


Figure 8.4: main window of GUI

To create a visualization, the user first has to select a time interval and then the buildings from and to which the movement should be visualized. The user has 2 options to select the time series for the current visualization:

1. Click on '1. Add dates' which will open the date picker dialog
2. Pick a day from the dropdown menu and click on '2. All days of week'

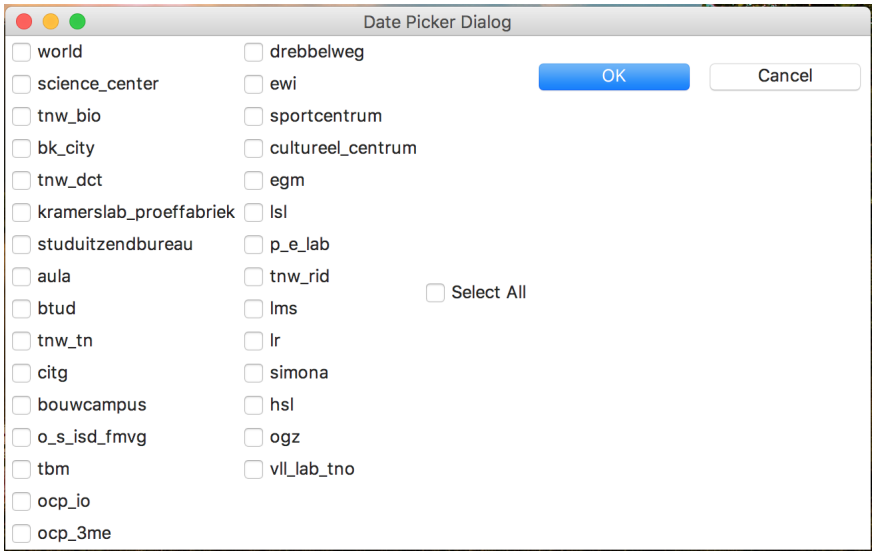


Figure 8.5: Buildings selection

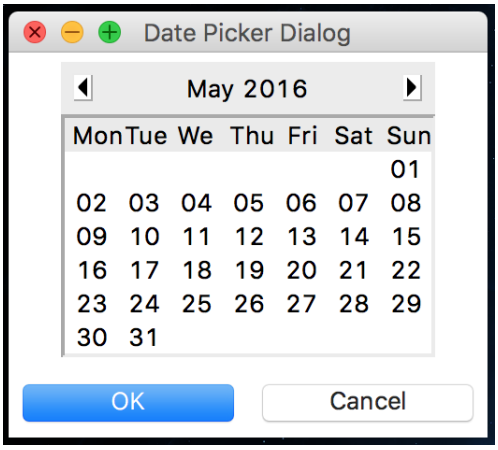


Figure 8.6: Dates selection

Option 1 can be used to select particular days without order. Option 2 can select 'every Tuesday' or even multiple recurring dates, such as 'every Monday to Friday'. It is also possible to combine option 1 and 2 to have for example 'every Monday and Friday the 13th of May'.

After selecting the time series, the user has to select the buildings from and to which the movement should be visualized. The 3rd and 4th button bring up the same dialog. This dialog shows checkboxes for every building. Every building that is check will be visualized. The user also has the option to select all buildings. If the user would like to see movement from and to the same building, the user can select the same buildings twice.

8.0.4. Maps

In order to get an overview about how people move on the campus and further more, find out movement patterns, a map visualization is essential. Map visualization consists of three parts:

1. base map: open street map is used as a base map. There are many labels on open street map, providing more context of the environment, so it is more clear and readable compared to other base maps like satellite images.

2. building markers: building markers show the locations of the buildings. Google maps marker style is used since it is commonly used in many map application. Because the shape of the building is not useful in analyzing movement patterns between buildings, each building is regarded as a point instead of a polygon, thus a node in the network,

3. lines: lines are the most essential part in map visualization, they represent movements between buildings.

In the first stage of map visualization, only base map and lines are taken into consideration, building markers are not shown on the map. The line width represents the amount of movement and movements are aggregated daily regardless of the timestamp of each movement during a day. This map visualization gives an overview of the movements over a day and between which buildings there are the most movements. The following maps show the difference of the amount of movement between April 11th (weekday) and April 17th (weekend).

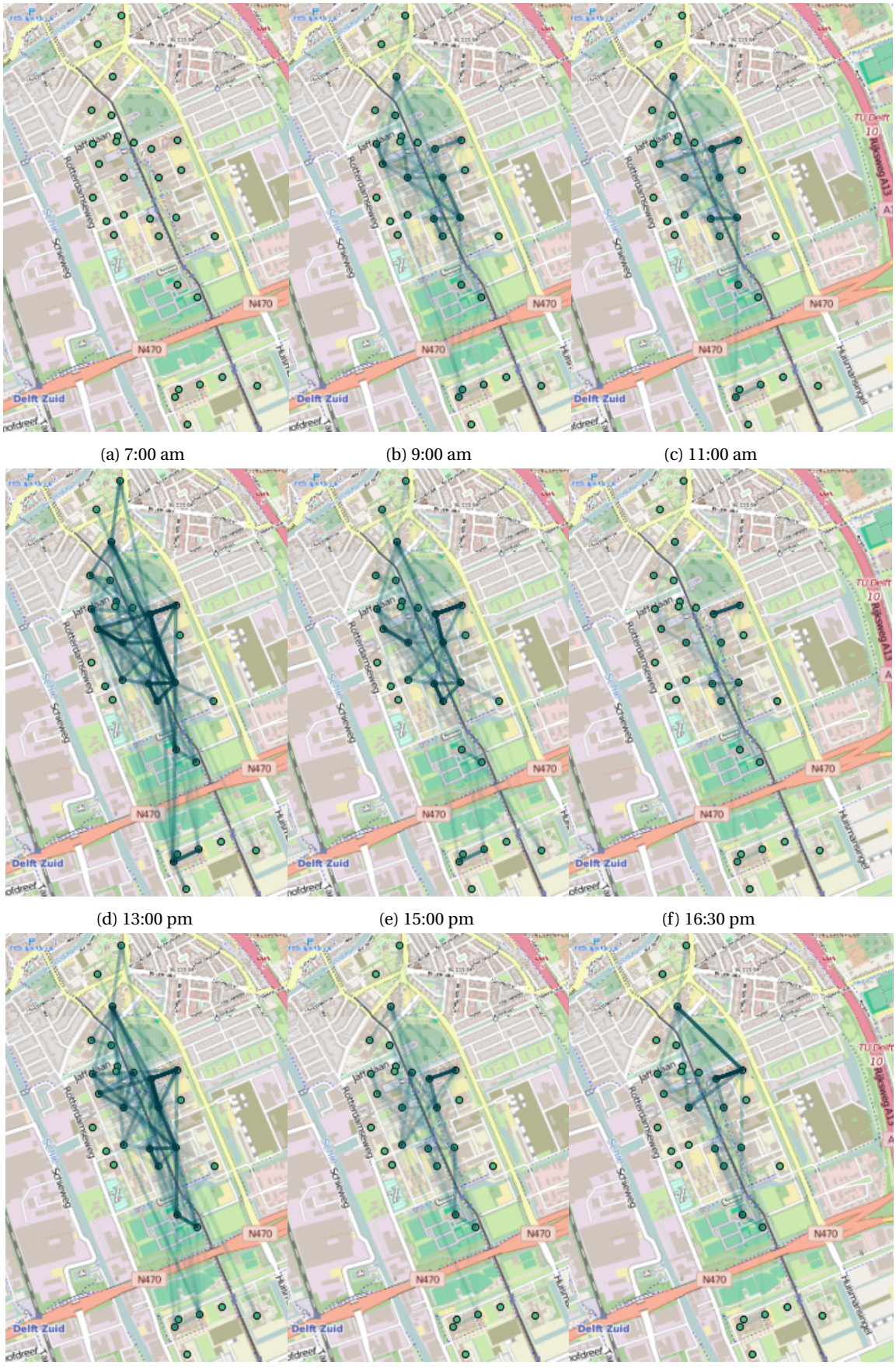
(a) April 11th, weekday

(b) April 17th, weekend

Figure 8.7: Static visualization

It's clear that between Aula and library, there are the most movements and the amount of movements is totally different on weekday and on weekend.

Given that movements are dynamic and occuring in both space and time, a dynamic map visualization is created to display individual movement over a day with temporal information. The following screenshots of the gif file show how the movements look like at a certain time of a day:



(a) 7:00 am (b) 9:00 am (c) 11:00 am

(d) 13:00 pm (e) 15:00 pm (f) 16:30 pm

(g) 18:00 pm (h) 20:00 pm (i) 22:00 pm

Figure 8.8: Dynamic visualization of movements, April 11th

In these pictures, the more movements there are, the less transparent the lines are. So generally speaking, from 7:00 am to 20:00 pm, there are two peaks at 13:00 pm and 18:00 pm. Hence, it is possible to get some insights about movement patterns from the animation. However, the dynamic map visualization doesn't provide detailed information to dig into but only an overview. So in order to find movement patterns, it is necessary to create maps containing more information, including time, direction and so forth.

Because the amount of data is big, it is more convenient to generate maps automatically so that it will fasten the progress of finding movement patterns. According to the three components of map, there is some information needed to be collected before visualizing movement on map. The locations of buildings are collected manually on Google earth based the campus map. These locations are exported as KML file and imported into QGIS. After adding geometry columns x and y, the csv file is created and imported into database. By using $ST\_MakePoint$ function, a geometry column is created in database. In summary, the building locations are stored as the structure described in following table:

id	name	geometry	x	y
0	world			
3	science_ center	0101000000042A7..	4.36939919846287	52.0072322181367
5	tnw_ bio	0101000000043AE..	4.37120211221402	52.0086132164098
8	bk_ city	0101000000077E3..	4.37053698152436	52.0056562098059
12	tnw_ dct	01010000007CA..	4.36891378927259	52.0040834950037
..	...	....	....	....

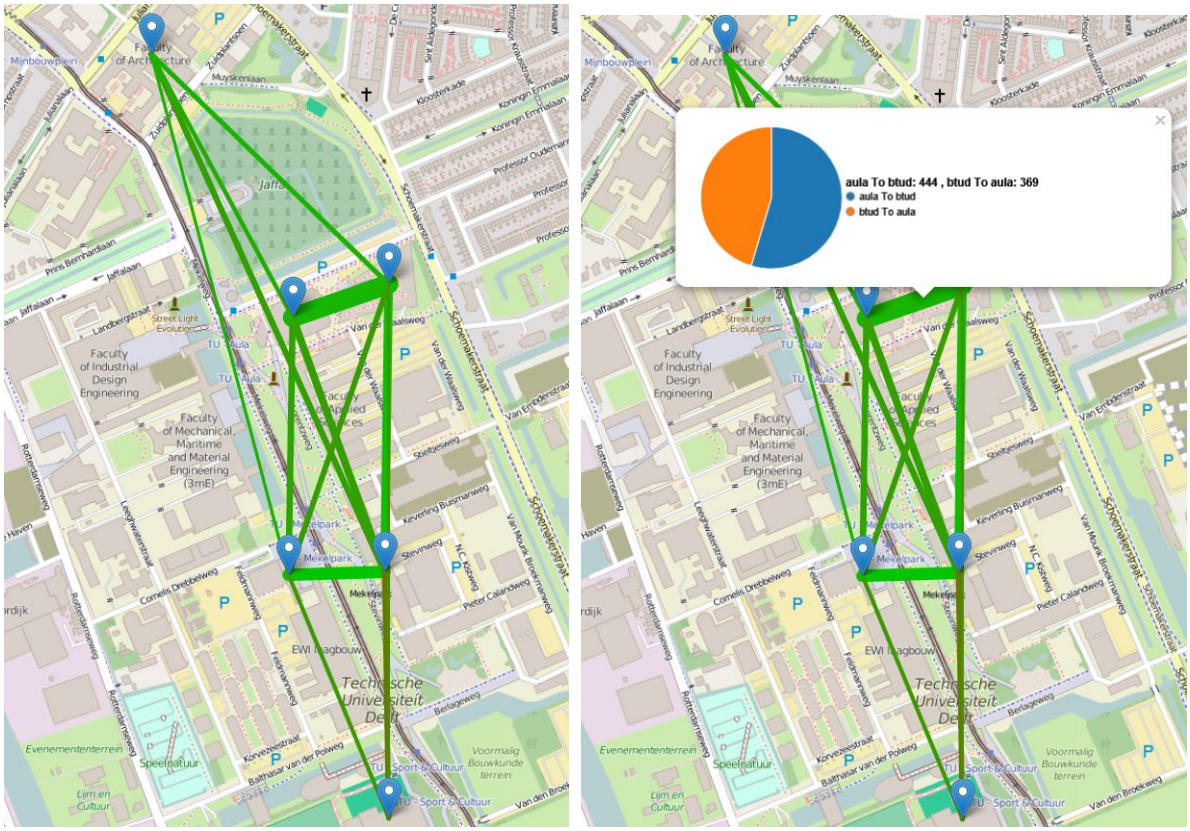
Table 8.1: Building data structure

There is a special 'building' called $world$ in the database. It is not an actual location, it is a virtual location which is used if someone is not scanned on the campus in a period of time. After storing the locations of buildings in the database, these locations will be extracted automatically from database to generate maps. There are two properties of lines used to deliver information:

1. width: line width is used to represent the amount of movements, but the amount is aggregated for both directions.

2. color: color is gradient from red to green. Red line means the movement is not symmetric that much more people move in one direction than the other, while green line means the movement is symmetric.

Based on this map visualization, users can choose certain dates and certain buildings to generate maps automatically. It makes it easier to find out movement patterns. Since not all buildings are chosen, the map will only display the movements between several buildings, which makes the map more readable:



(a) Amount of movements on April 25th

(b) Amount of movements in pie chart

As shown in the map, the lines are in different colors, which shows the symmetry of the movements. If the user is willing to know more about the movement, it is also possible to click on the line to check the amount of the movements for each direction in detail, and there will be a pie chart showing how symmetric the movements are. With the map visualization, it is easy to focus on movements which are special or interesting.

8.1. Introduction

8.2. Theory / methods

8.3. Implementation

8.4. Results

The movement trajectories and the amount of movement between buildings can be visualized in maps and bar charts. The previous sections explained how the data is transformed and this section focusses on the results that can be derived from this data. Bart Valks and Iljoesja Berdrowski stated some questions that arise in their line of work and this section will try to answer these questions with the visualization in both maps and bar charts.

8.4.1. Movement to the Aula on weekdays

The department of FMRE would like to know if the faculty of Applied Sciences uses the restaurant facilities of the Aula more than other buildings, due to the fact that the two buildings are connected with a bridge on the first floor.

The graph below shows the average movement of people to the Aula on weekdays. Clearly a peak can be distinguished in the morning between 8:00 and 9:00, around lunch time and in the afternoon between 17:00 and 18:00. The morning and afternoon movements represent people moving from home to the aula and back home.

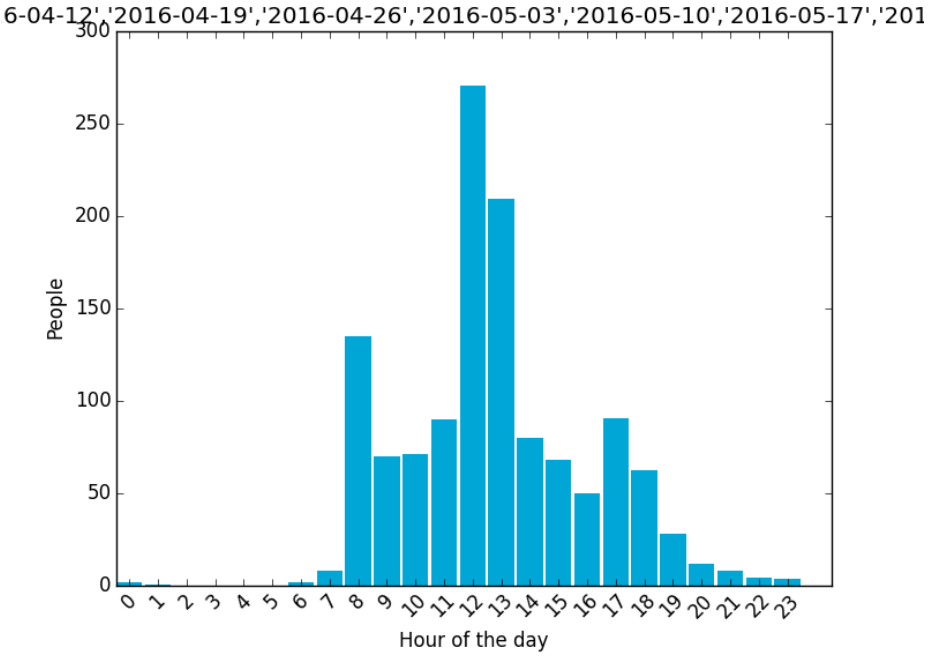


Figure 8.10: All to Aula

The graph however, says nothing about which buildings contribute the most to the movement to the aula. The map image below shows the top 10 buildings with movement to the aula. It is clear that most of the movement comes from the Library. But if leaving the Library out of the equation, it is clear to conclude that the faculty of Applied Sciences uses the aula more than other faculties. The movement from TNW to the aula is 5000 people over the whole dataset, where other faculties don't get higher amounts than 2500.

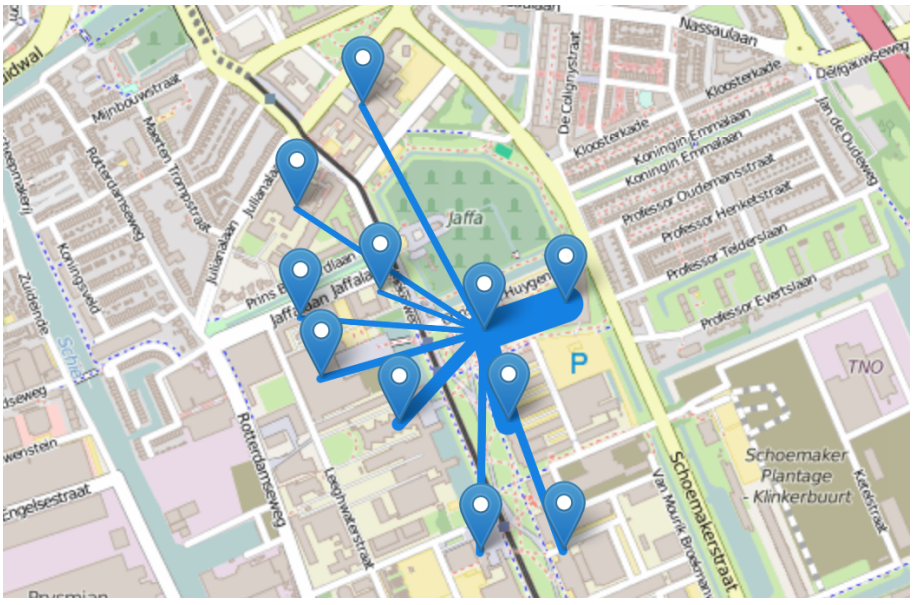


Figure 8.11: Maps from all buildings to Aula

This partly confirms the assumption that FMRE made, but to be sure, the movement from the faculty of Applied Sciences must also be checked, in order to see if the movement to the aula is no exception. The result of this visualization is shown in the map image below.

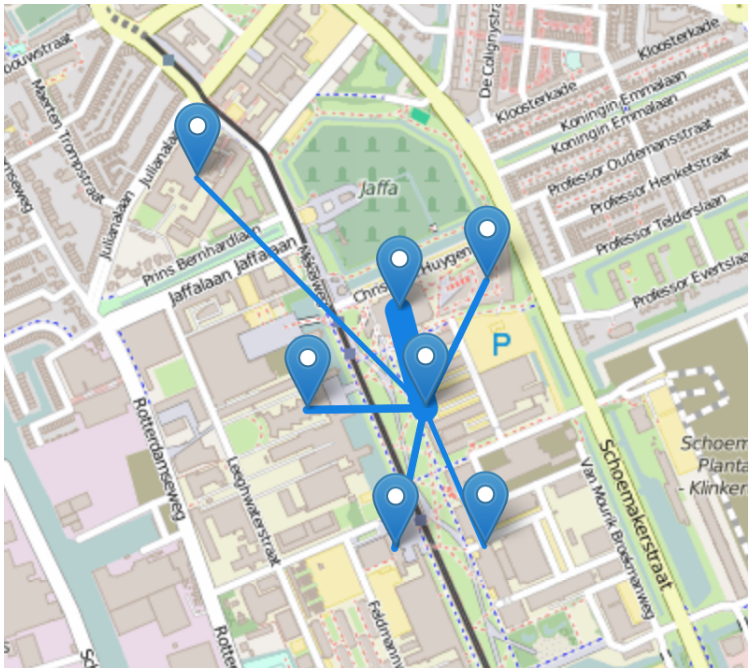
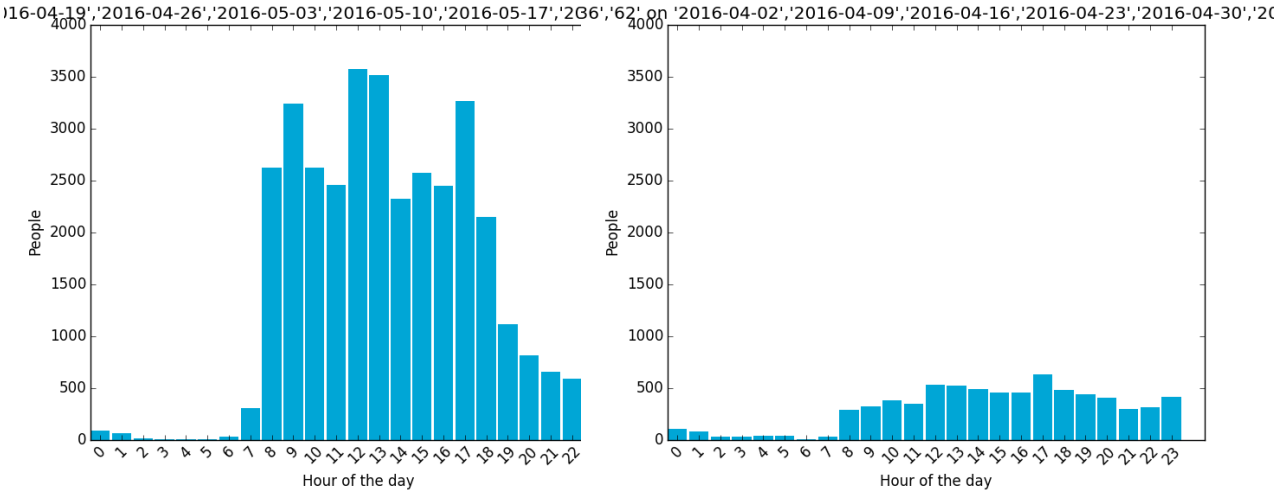


Figure 8.12: Maps from all buildings to TNW

This image shows that indeed most of the movement originating from the faculty of Applied Sciences is going to the Aula. The bar chart provides more insight in when this movement is taking place, which is during lunch, as expected.

8.4.2. Movement on weekdays vs. weekends



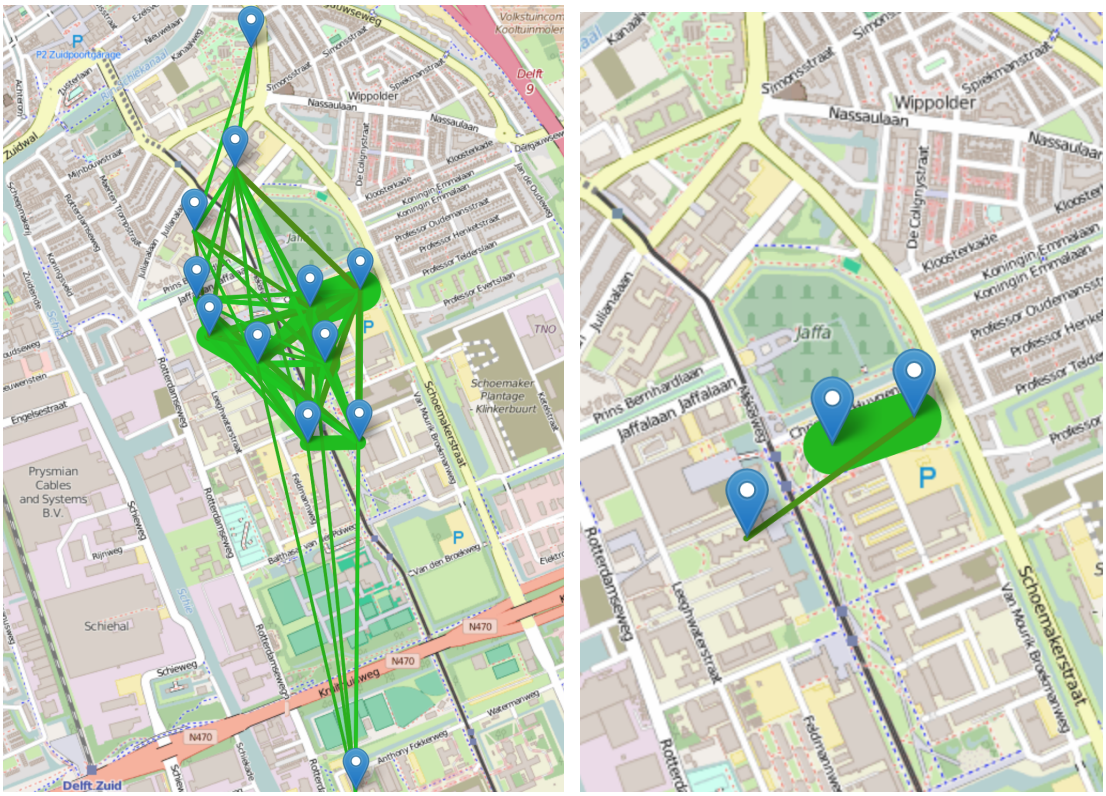
(a) Movements on weekdays

(b) Movement on weekends

Figure 8.13: Bar charts of the movements

The figures above show the movement from and to the 12 most used buildings on campus. Figure 8.13a shows the barplot for the weekdays, Figure 8.13b shows the barplot for the weekends. It is clear to see that during weekends, there is a lot less movement during weekends. Especially during lunchtime, we can see a peak during weekdays and in the morning and afternoon. In the weekend the movement to the faculties is apparently

much less and more spread out over the day.
Interesting to see is the movement in the early morning, between 00:00 and 4:00. The library is only open from 8:00 to 2:00, but the bar chart alone does not provide enough information to draw conclusions about these movements.



(a) Movements on weekdays

(b) Movements on weekends

Figure 8.14: Maps of the movements

Figure 8.14a and Figure 8.14b show the spread of the movement over the whole campus. Now it becomes clear that movement during weekdays is spread out over all faculties, but during weekends is only focused on the Library. There is however one exception, the faculty of 3ME. This could be explained by staff using the building with their campus card.

It is interesting to see that during weekdays the movement from and to a faculty is almost always symmetric, whereas in weekends this is certainly not the case.

8.4.3. Architecture as an island

The department of FMRE also assume that architecture students and staff have the tendency to stay in their faculty and move less to other buildings on campus than other faculties.
Their question can be answered when looking at the movement between the faculty of architecture and all other buildings. This shows the amount of movement to other faculties, but these amounts need to be compared to the movement from other faculties (for this question IO, CiTG and LR are considered) to other buildings.

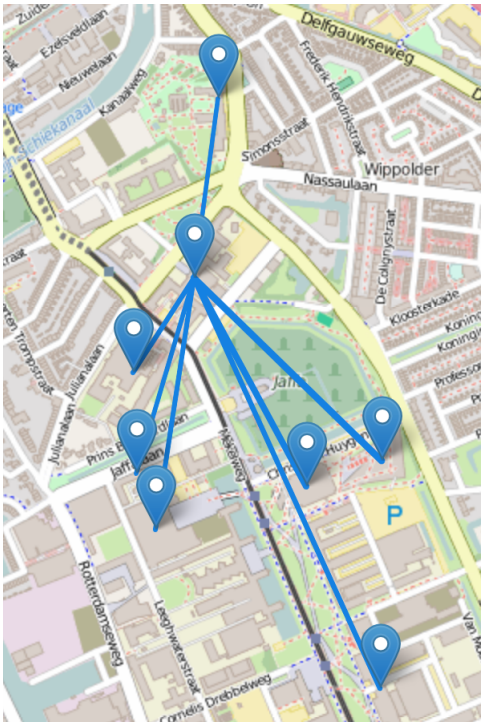


Figure 8.15: Movement from the faculty of Architecture

Figure 8.15 shows the movement between the faculty of architecture and other faculties. The map shows the 7 most used buildings, the movement to other faculties is only 2% of the total movement and is left out for readability. The total amount of movement from Architecture is 4.239 people. The largest movement to any other faculty is to the Library with 1.043 people.

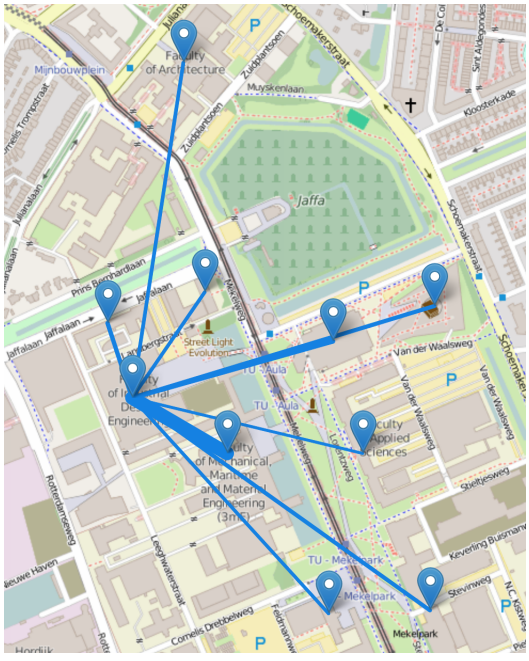


Figure 8.16: Movement from the faculty of Industrial Design

Figure 8.16 shows the movement from the faculty of Industrial Design to other faculties. The total amount of movement from IO is 10.933 people and the biggest movement is to the faculty of 3ME, with 4.927 people.

This is already a lot more movement than the faculty of architecture.

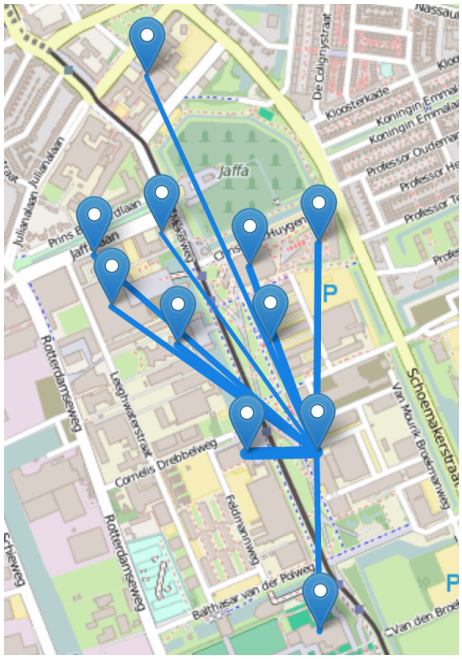


Figure 8.17: Movement from the faculty of Civil Engineering

Figure 8.17 shows the movement from the faculty of Civil Engineering to other faculties. The total amount of movement from CiTG is 11.035 people and the biggest movement is to the faculty of EWI, with 2.897 people. This is even more movement than IO.

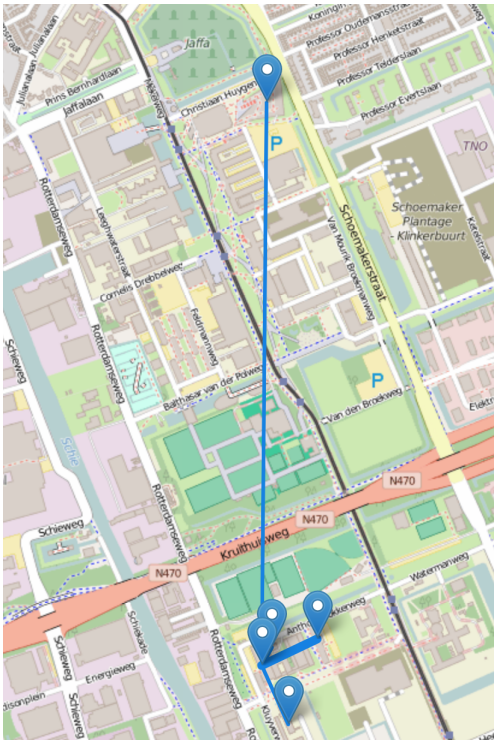


Figure 8.18: Movement from the faculty of Aerospace Engineering

Figure 8.18 shows the movement from the faculty of Aerospace Engineering to other faculties. The total amount of movement from AE is 4.348 people and the biggest movement is to the Fellowship, with 2.435 people.

Looking at these figures we can clearly see that the movement from the faculty of architecture is much less than the movement from Civil Engineering and Industrial Design. However, the faculty of Aerospace Engineering seems to be even more isolated than architecture.

8.4.4. Movement from and to the campus

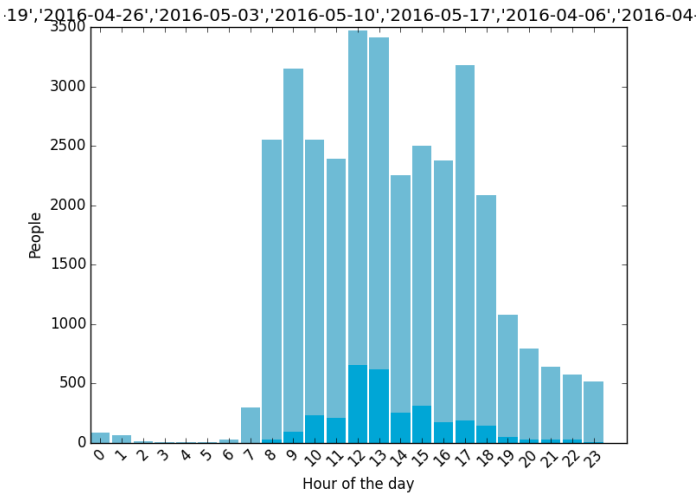


Figure 8.19: Movement from and to the campus

The aggregated movement from the different faculties is shown in Figure 8.19. For this bar chart, a distinction is made between movement that is either from or to the campus(in light-blue) and movement between buildings on campus (darker blue). This will result in a more accurate visualization of the data.

However, the movement from or to the campus is derived from finding gaps of more than one hour in the data, because devices going offline for more than one hour could be considered moving away from the campus. But this does also include devices that are for some reason turned off for more than one hour, such as laptops during lunch breaks or lectures. For these graphs to really accurately show the movement from and to the campus, devices should be categorized into dynamic devices, such as mobile phones and tablets, and static devices, such as laptops. That way, only the dynamic devices can be considered and the graph would be more realistic.

9

Trajectory patterns

9.1. Introduction
9.2. Theory / methods
9.3. Implementation
9.4. Results

This GSP attempts to identify people's movement patterns from anonymized wifilogs. To solve this problem, individual trajectories must be discovered. 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 obeservation 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 positioining method. The size of each Wi-Fi cell determines the spatial resolution, as the location of mobile devices is estimated at the origin of a AP. The size of each Wi-Fi cell depends on the distribution of APs. 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.

First, this chapter will describe the extraction of locations of a user. Then the mining of individual trajectories from a anonymized Wi-Fi scanlist is described. Subsequently, the mining of movement patterns in time or space is described.

9.4.1. Location extraction

A location represents a geographic position where a user stays. 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 WiFi 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.

9.4.2. 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:

$$p1 \rightarrow p2 \rightarrow p3 \rightarrow \ldots \rightarrow pn$$

Given a time interval, there is a set of individual trajectories S = {t1, t2, t3,...,tn} where each ti is the trajectory over a time interval of one user.

9.4.3. Trajectory Pattern

From a set S of trajectories, different patterns can be identified using seqeuntial pattern mining algorithms. Frequency of a trajectory by all users of the campus can be detected. This can be represented as a trajactoy 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 lenght of a trajectory can be discovered. This allows for identification of movement patterns of a specific lenght n. Also, when location is not considered, but only the lenght of a trajectory, the mobility pattern of an individual can be discribed in terms of how many times he/she re-locates.

10

Indoor spatio-temporal movement patterns

11

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.

12

Recommendations

Recommendation for future research is first of all the filtered of static devices, as this could improve the quality of the results. Furthermore additional research is required to find methods that can be used for determining the quality of the results, especially concerning the movement from and to the campus.

The detail of the data could be improved by increasing the frequency with which the eduroam system is scanning. This could especially support analysis on the use of different entrances. With the present frequency of one scan round per 5 minutes it is highly likely that the first scan when someone is entering a building is not at the entrance. Detail on which routes are travelled between buildings and validating movement from and to the campus could possible be accomplished by strategically placed scanners outdoor. Finally it is recommended to store the locations of access points digitally in a single map. This would support the process of identification of movement patterns inside buildings.

Finally caution should be taken with the dataset regarding people their privacy. It is relatively easy to identify a particular user and track that person via the data.

12.1. Entrances and exits

This section will describe the undergoing process in order to know how frequent the entrances and exits of a building are used. Knowing this will give insight into the use of a building, the spatial context and the relation between these two. Our hypothesis is that access points located near the entrance(s) of a building are most frequently used as first access point when entering a building, and as last access point when leaving a building. Firstly, an approach will be presented that does not take in account that devices might get scanned when passing by the building. In the second approach we will make use of the pre-processed data which excludes the devices that get scanned when passing by the building.

12.1.1. First approach:including devices passing by

The first approach makes use of the raw wifilog data, by finding the part in a sequence in which a device is scanned by an access point in a building and is subsequently scanned in another building. With the location of the access points known, we hope to get insight into the use of an entrance or exit location in a building.

53	08-BK-City	A-08-F-001	6724
54	08-BK-City	A-08-G-103	2710
55	08-BK-City	A-08-E-008	2447
56	08-BK-City	A-08-G-004	2280
57	08-BK-City	A-08-B-007	1922
58	08-BK-City	A-08-B-003	686
59	08-BK-City	A-08-G-101	563
60	08-BK-City	A-08-H-001	533
61	08-BK-City	A-08--E-003	484
62	08-BK-City	A-08-H-004	338
63	08-BK-City	A-08-J-005	335
64	08-BK-City	A-08-D-005	294
65	08-BK-City	A-08-H-003	286
66	08-BK-City	A-08-D-010	276
67	08-BK-City	A-08-F-103	251
68	08-BK-City	A-08-E-001	202
69	08-BK-City	A-08-E-102	201
70	08-BK-City	A-08-C-003	197
71	08-BK-City	A-08-D-102	180
72	08-BK-City	A-08-D-004	174
73	08-BK-City	A-08-J-007	167
74	08-BK-City	A-08-H-002	163
75	08-BK-City	A-08-E-005	151
76	08-BK-City	A-08-B-101	133
77	08-BK-City	A-08-B-005	130

Figure 12.1: A segment of the resulting table after querying

The stays in which the device is scanned once are not filtered out. These single scans imply that a person with the device only passed by the building, thus was not really located in the building.

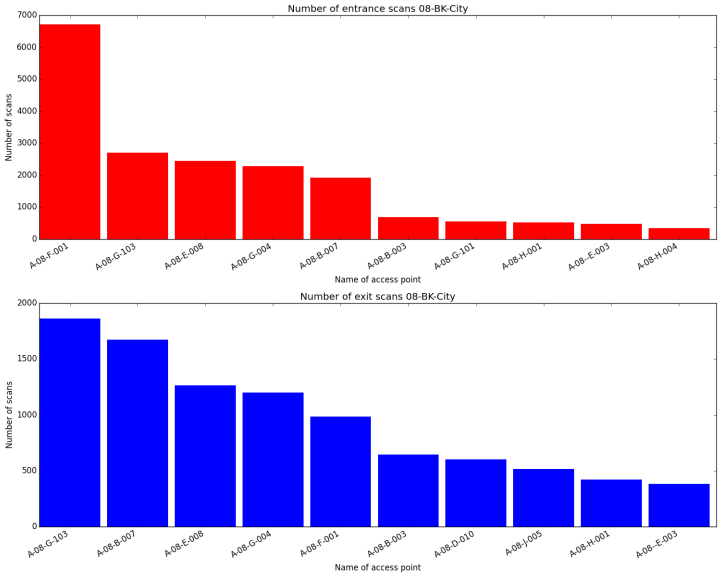


Figure 12.2: Most frequently used entrance and exit access points in BK-City

In order to know whether these access points(see Figure 12.2) are located near an entrance, the access point maps of BK-City is used. The access point maps are the building plans enriched with the location of each wifi access point installed in the building. Currently, the access point maps of BK-City are the only ones available. Looking at the location of the access points with the highest frequency, gives an interesting result (Figure 12.3).

Figure 12.3: The location of the most frequently used entrance and exit access points

Most of the frequently used access points are located at the western part of BK-City (Figure 12.3). Also, there is no entrance or exit located near most of these access points. Knowing that lots of people are passing in the street next to the western part of the building, we can conclude the result of this analysis is distorted due not filtering out the devices that get scanned when passing by the building.

12.1.2. Second approach:excluding devices passing by

Figure 12.4 depicts the table as a result of the pre-processing as described in chapter 7. The records represent the stays for each mac, including the first and last access points (ap_ start and ap_ end).

	mac text	building integer	ts timestamp without time zone	te timestamp without time zone	ap_start text	ap_end text
1	/4SpD2kESBS	0	2016-03-30 23:34:53	2016-04-25 09:15:50	NULL	NULL
2	/4SpD2kESBS	32	2016-04-25 09:15:50	2016-04-25 19:11:49	A-132-0-030	A-132-0-073
3	/4SpD2kESBS	0	2016-04-25 19:16:50	2016-05-13 13:10:56.946	NULL	NULL
4	gLrQi1Ex+z8	0	2016-03-30 23:34:53	2016-04-05 13:19:13	NULL	NULL
5	gLrQi1Ex+z8	36	2016-04-05 13:19:13	2016-04-05 14:16:11	A-36-0-024	A-36-0-024
6	gLrQi1Ex+z8	36	2016-04-05 14:37:15	2016-04-05 15:18:46	A-36-0-024	A-36-0-024
7	gLrQi1Ex+z8	0	2016-04-05 16:26:16	2016-04-26 14:07:39	NULL	NULL
8	gLrQi1Ex+z8	36	2016-04-05 15:44:59	2016-04-05 16:26:16	A-36-0-058	A-36-0-027
9	gLrQi1Ex+z8	0	2016-04-26 15:48:05	2016-05-03 13:29:25	NULL	NULL
10	gLrQi1Ex+z8	21	2016-04-26 14:07:39	2016-04-26 15:48:05	A-21-0-015	A-21-0-061
11	gLrQi1Ex+z8	0	2016-05-03 13:39:51	2016-05-13 13:10:56.946	NULL	NULL
12	NH3C//jQ5kw	0	2016-03-30 23:34:53	2016-04-04 15:57:52	NULL	NULL
13	NH3C//jQ5kw	21	2016-04-04 16:03:06	2016-04-04 17:52:00	A-21-0-048	A-21-0-048
14	NH3C//jQ5kw	0	2016-04-04 18:07:11	2016-04-05 08:39:54	NULL	NULL
15	NH3C//jQ5kw	32	2016-04-05 08:39:54	2016-04-05 12:22:08	A-132-0-048	A-132-0-048
16	NH3C//jQ5kw	32	2016-04-05 12:48:07	2016-04-05 15:24:10	A-132-0-047	A-132-0-033
17	NH3C//jQ5kw	0	2016-04-05 15:39:34	2016-04-06 11:23:31	NULL	NULL

Figure 12.4: A segment of the table as a result of the pre-processing

The table also includes 'world' (in the Figure 12.4 represented by NULL) which implies the device is not located on the campus.

The following simple SQL statement is used to plots the most frequently used entrance access points.

```
SELECT ap_start, count(*)
FROM table
GROUP BY ap_start
ORDER BY count desc;
```

Ap_ end is used, instead of ap_ start, for plotting the most frequently used exit access points in a building.

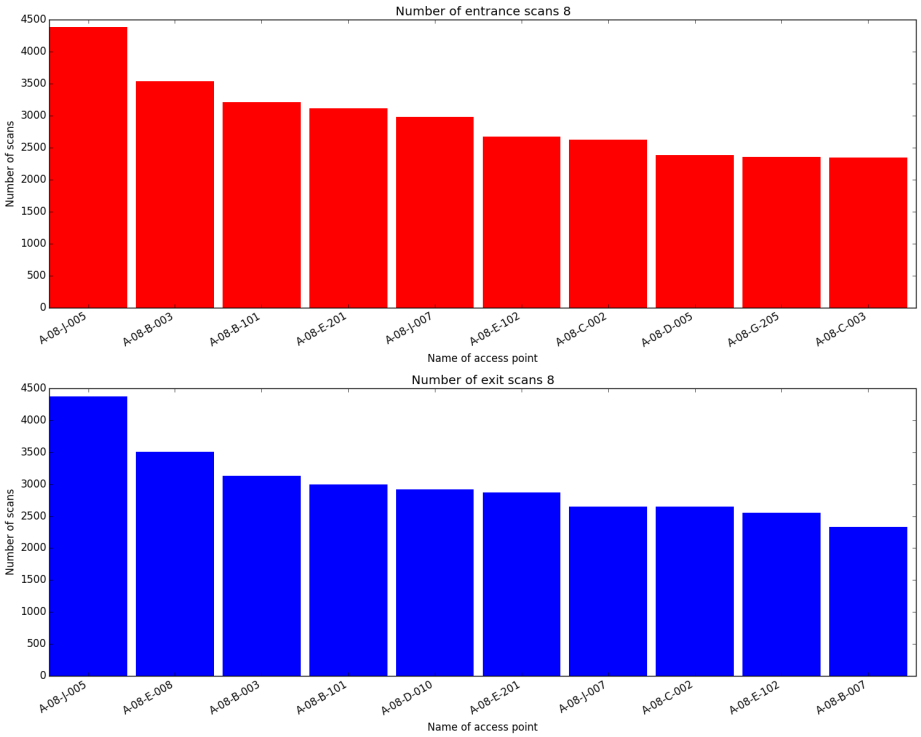


Figure 12.5: The most frequently used entrance and exit access point for BK-City

The most frequently used access point, A-08-J-005, is not located near an entrance or exit (see Figure 12.5). This is different than expected. Although it is not very logical in the first place, it still might be one of the first or last access points a device connects with. The reason for this is that the A-08-J-005 access point is placed in in an open space without many objects that could block the wifi signals.

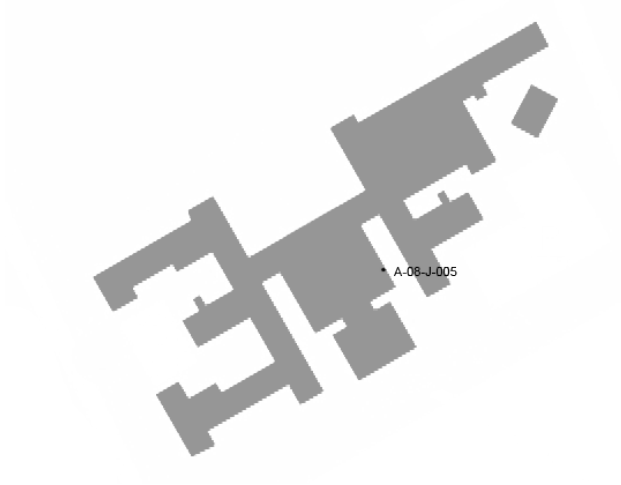


Figure 12.6: The location of the most frequently used entrance and exit access point, according to our second approach

We expected the access point to be located much closer to an entrance or exit. The plan is to set up an experiment in order to justify the unexpected result. In this experiment we will check to what access points different devices (laptops and mobile phones) connect when entering or leaving a building.

12.1.3. Frequency of entrance and exit access points

This section will describe the analysis on the frequency of entrance and exit access points. As described in subsection 12.1.2, the most frequently used entrance and exit access points are not always are located near

an entrance or exit. Though it is still possible to analyze how frequent these access point are used. The results will be aggregated, meaning it represents more than a single day.

Entering First we will take a closer look at an access point which appears to be one of the first that scans the device. This will be A-08-J-005 in BK-City, see Figure 12.5 in subsection 12.1.2. The chart below shows the frequency of entrance access point A-08-J-005 for devices entering BK-city, over a 24 hour time period.

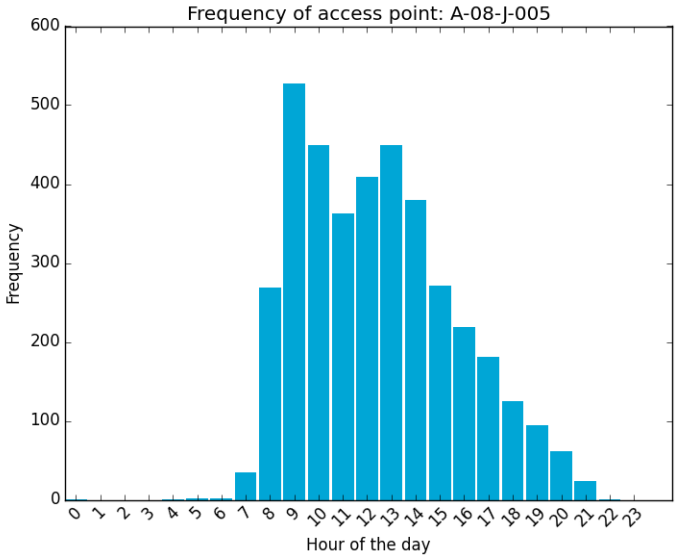


Figure 12.7: Frequency of entrance access point A-08-J-005

The chart shows two peaks; in the morning and around 12pm to 1pm. This is in line with what we expected. In the morning a large group enters the building and around 12pm to 1 pm a large group enters the building after the lunch break.

Exiting

For the exit situation, again the A-08-J-005 access point will be used. This access point also appears to be the most frequently used exit access points. The chart below depicts the frequency of devices leaving the building over a 24 hour period.

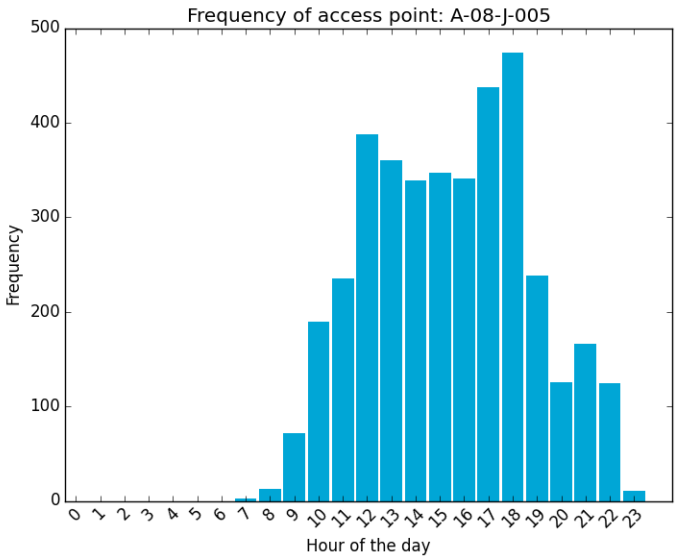


Figure 12.8: Frequency of entrance access point A-08-J-005

The chart shows three interesting peaks, which is in line with what we expected. The first one, around 12am to 1pm, is due to people that leave the building, most probably for lunch. The second peak, around 6pm to 7pm, is due to people that go home for diner. The last one, around 9pm to 10pm, is due to the closing time of the building.

12.2. Association rules

When a trajectory is simplified into a set of distinct buildings that the person visited, association rules for buildings can be derived. In this case the rule describes the set of buildings, or buildingset, that are commonly visited in combination. For example the rule {BK_City, Aula} => {Library} tells that a group of people who visited the buildings BK_City and Aula also visited the Library.

As association rule mining does not consider the order of buildings, nor the time spent in a building, it is important that these variables are appropriately handled and noise is filtered out prior running the algorithm.

In the first version the buildingsets were stored in a table as below, where the field *mac* contains the mac-address of a device and each remaining field represents a building. Value 1 is given if the device was recorded in a building, otherwise no value is given. This binary encoding is rather simplistic as it does not consider the amount of time spent in a building and therefore it does not allow to differentiate between occasional or regular visits.

mac	aula	bk_city	bouwcampus	btud	ctig	...
A	1	1			1	
B			1	1		
C		1			1	
D	1					
E	1		1			

Table 12.1: uncategorized buildingset table

Therefore in the second version a distinction between *occasional, regular* and *frequent* stays was added to the buildingsets. The division between the categories is based on the 40 hour workweek and 1.5 hour lecture durations (see Table 12.2).

Category	hours/week	ID
occasional	≤ 0.5	1
regular	$> 0.5, \leq 5$	2
frequent	> 5	3

Table 12.2: Stay duration categories

The trajectories of approximately 14,000 devices were used to create the first set of association rules with categorized stay duration. At this stage only the noise was filtered from the data but not the stationary devices, and people carrying two devices were not accounted for. The time range of trajectories spanned from 31.03.2016 to 02.05.2016, approximately one month.

Although there are several measures to evaluate the interestingness of an association rule (Zhang et al. 2009), only *support* and *confidence* were used for testing purposes.

Support "The support for a rule is defined to be the fraction of transaction in the dataset that satisfy the union of items in the consequent and antecedent of the rule." (Agrawal, Imieliński, and Swami 1993). In case of the rule {BK_City, Aula} => {Library}, the support is the percentage of the total dataset that includes BK_City, Aula and Library.

Confidence Confidence measures the strength of the rule, and is considered as a conditional probability. In case of the rule {BK_ City, Aula} => {Library}, the confidence is the probability that Library is in the trajectory if both BK_ City and Aula are in the trajectory (Agrawal, Imieliński, and Swami 1993; Anbukkarasy and Sairam 2013).

The most interesting rules are displayed in Figure 12.9:

Supp ▲	Conf	Covr	Strg	Lift	Levr	Antecedent		Consequent
0.02	0.86	0.02	6.92	6.51	0.01	drebbelweg=2, ewi_lb=2	→	ewi_hb=2
0.01	0.74	0.01	24.80	3.38	0.00	btud=2, drebbelweg=2, tbm=2	→	citg=2
0.01	0.70	0.01	30.21	3.21	0.00	aula=2, lr=2, ocp_io=2	→	citg=2
0.01	0.70	0.01	29.22	2.44	0.00	aula=2, lr=2, ogz=2	→	btud=2
0.01	0.72	0.02	6.60	5.49	0.01	btud=2, ewi_lb=2	→	ewi_hb=2
0.01	0.73	0.01	15.17	5.52	0.01	aula=2, btud=2, ewi_lb=2	→	ewi_hb=2
0.01	0.72	0.01	16.24	5.44	0.00	btud=2, citg=2, ewi_lb=2	→	ewi_hb=2
0.01	0.90	0.01	17.31	6.81	0.01	btud=2, drebbelweg=2, ewi_lb=2	→	ewi_hb=2
0.01	0.85	0.01	20.16	6.43	0.00	citg=2, drebbelweg=2, ewi_lb=2	→	ewi_hb=2
0.01	0.74	0.01	10.08	5.59	0.01	ewi_lb=2, tnw_tn=2	→	ewi_hb=2
0.01	0.86	0.01	21.85	6.51	0.00	drebbelweg=2, ewi_lb=2, tnw_tn=2	→	ewi_hb=2
0.01	0.72	0.01	25.66	2.52	0.00	aula=1, tbm=2	→	btud=2

Figure 12.9: Building set

In the buildingset of approx. 14,000 devices 2% was recorded in all of the buildings *Drebbelweg, EWI-LB, EWI-HB* (Support = 0.02). There is an 86% chance that if a device is recorded in the buildings *Drebbelweg, EWI-LB*, then it is also recorded in *EWI-HB* (Confidence = 0.86). And they spent on average between half hour to five hours a week in each building (drebbelweg=2, ewi_ lb=2, ewi_ hb=2).

12.3. Distinguishing user groups

12.4. Occupancy

12.5. AP system

12.6. Data reasoning

12.7. Visual exploration

13

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.

14

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.