A Methodology for Deriving Aggregate Social Tie Strengths from Mobility Traces

By
Tristan Brugman

Master's Thesis for the master Computer Science, specialization Data Science and Smart Services

Completed at

University	Twente University
Faculty	Faculty for Electrical Engineering, Mathematics and Computer Science
Chair	Design and Analysis of Communication Systems (DACS)
Date	March 2017

Author

Name	Tristan Brugman
E-mail	t.w.r.m.brugman@student.utwente.nl

Graduation Committee

Dr. Geert Heijenk
Dr. Mitra Baratchi
Prof. Dr. Ir Maarten van Steen

Thesis supervisor: Geert Heijenk Tristan Brugman

Abstract

The degree of social connectedness of people in a location has a large impact on how that place functions, and often influences our decision whether or not to visit it. Similarly, knowledge of a location's social connectedness could enable a variety of important applications, such as improved elderly care and socially aware smart phone applications. These and other applications would benefit from automatically discovering the social characteristics of the place they are in, but this information is not always obtainable. Previous studies have devised methods to infer the social tie strengths of visitors from information specific to certain communication protocols, such as Wi-Fi, but they cannot be used by devices that do not use these protocols.

In this thesis, we propose a novel method to infer aggregate social tie strengths from device mobility data. The main benefit of this method is its general applicability, as it could be used by any application that has knowledge of devices entering and leaving the specified location. The method works by training a regressor on a subset of a large collection of mobility features and known social tie strengths. Then, this regressor predicts the social tie strengths for devices present in the location at a given moment in time, and outputs an aggregate score.

In order to evaluate the method, we tested it on a real-world data set gathered by Wi-Fi sensors for several months. We found that the accuracy of the proposed method highly outperforms that of a state-of-the-art baseline methodology we based on a recent study. Additionally, we tested the proposed method on several modifications of the real-world data set, in order to simulate more difficult environments. On these data sets, too, the method maintains a high accuracy, signifying its robustness.

Acknowledgments

This thesis is the result of the final project of the master Computer Science at the University of Twente. It was made possible by the help of several people, who I would like to thank here.

First of all, I would like to thank my daily supervisor Mitra Baratchi, who has helped me on numerous occasions during the project. Her valuable feedback and assistance have greatly helped me both in performing the research and writing this thesis. This work would not have been possible without her.

I would also like to thank Geert Heijenk for being part of my graduation committee, and introducing me to Mitra. I believe his feedback at the end of the project helped improve this work's scientific value.

Additionally, I would like to thank Maarten van Steen, who is also part of my graduation committee. He helped shape the direction of this research by giving valuable insights on its social aspects at its start, and also provided very useful scientific feedback at its end.

Finally, I would like to thank my friends and family. They have supported me during the course of my whole study, and their curiosity about this research encouraged me as well.

I hope you enjoy reading this thesis.

Tristan Brugman

Contents

1
Introduction

In our everyday life, we often want to know how social a place is before visiting it: if we want to meet new acquaintances, we may visit a pub, but if we want to study or read, a location like a library is more appropriate. Likewise, many software applications would benefit if they had knowledge about the social characteristics of places, but this information is not easily obtained from the environment. So, in order to give many devices some degree of social awareness, we need to find a method that infers social connectedness from information that *is* generally available. The study presented in this thesis aims to solve that problem.

Specifically, we set out a methodology to extract the social connectedness of a location, based on mobility features alone. The main output of the proposed methodology is a score that indicates the aggregate social tie strength of the people present in the location at a certain moment in time. This methodology enables both visitors and location owners to automatically learn about the social connectedness of visitors, while preserving the privacy of those visitors as much as possible.

1.1 Applications

The main results of the method that will be created during the research are the pair-wise social tie strengths between devices and a value for the location's aggregate social connectedness. Both the social network and the aggregate score have many important applications. For example, they could be used to inform user applications of the type of location they are present in, to inform the policies at care facilities, or to optimize mobile ad hoc networks. In this section, we describe a number of applications of the proposed method.

1.1.1 Improving Care

It has long been known that social ties in a community tend to positively impact the health of its members. Multiple studies from the 1970s onward have established the positive effects of social integration on psychological health [43, 25, 24]. More recent studies have established that social relationships are beneficial for physical health, as well [8, 19]. The impact of social ties on health is becoming especially critical with the rapid aging of the human population [38], as previous studies have found that social isolation is a significant risk factor for the mortality of the elderly [40, 20].

Residents of facilities such as nursing homes, retirement homes and retirement communities could gain better care if the facilities were better aware of their social connections. For example, based on the absence of strong social relationships between residents, facilities could modify their policies to include more social activities. Social connectivity could also be used as a statistic for governmental oversight, and could improve care on a national level. If social connectivity could be determined automatically, the process could be performed more consistently, and for a lower cost to the facility.

1.1.2 Social User Applications

Nowadays, many user applications have a social component, and they may benefit from automatically discovering that people with a social connection are currently present in the same location. Examples include smart phone applications like Facebook and computer programs like Steam. After discovering its owner is in a social location, these applications could suggest social activities, such as sharing messages or making payments, and participating in a multiplayer video game.

More generally, a measurement of the social connectedness over a period of time could be used to create a social fingerprint of a location, which may be used to identify the type of location or to distinguish it from other locations. It could also be used to improve a more general location fingerprinting method, as in [3].

1.1.3 MANETs

Mobile ad hoc networks (MANETs) allow nearby devices to communicate with each other or with a larger network, without the need of a larger infrastructure to do so. This allows devices to access the Internet without a direct connection to it, as other nearby devices can act as routers, relaying information between the device and the Internet. It can also increase the data confidentiality and decrease the required resources and delay for communication between nearby devices, as the larger Internet can be bypassed.

In MANETs, each device needs to select some routers from the available nearby devices. In this selection process, devices that stay in range longer should be preferred over devices that are expected to be inaccessible. Since people with stronger social relations tend to meet more often and stay in the same location longer, the existence of a social link could be used to optimize this process. Similarly, the algorithm may perform differently based on the current nearby social connectedness. Previous studies have already used social information in order to improve the forwarding of data [31, 21].

1.1.4 Scientific Research

Finally, knowledge of social connectedness or composition may be useful scientific information, both for biology and the social sciences. In both sciences, it may be used to gain insight into group and social bonds formation [46], and in- and out-group behaviors. Instead of probe data, mobility data may be gained by using video cameras or (in the cases of tracking animals,) tracking chips. It would also be useful in the field of epidemiology, as understanding social networks is key to understanding how diseases spread [5, 12].

1.2 Problem Statement

The goal of this study is to use mobility information from visitors' devices from a single location to derive the aggregate social connectedness for that location for a given moment in time. More formally, we can define this as follows. Given a location, we consider detections of the presence of devices in that location. Each detection is a tuple <d, t>, in which *d* represents the visiting device and *t* represents the moment in time that the device is detected. We define a mobility trace as a collection of detections. Additionally, we consider the strength of the social relation between the owners of devices: each pairwise social tie strength is a score *s* for each tuple <d1, d2>, in which *d1* and *d2* represent different devices and *s* represents the strength of the social relationship between the owners of *d1* and *d2*. Finally, we define the aggregate social tie strength as the social tie strength between a group of users: it is a score *s* for each set {d1, d2, d... }, in which each *d* represents a different device. Given a mobility trace and the pairwise social tie strengths for some pairs of devices in the mobility trace, we are interested in finding a methodology that infers the aggregate social tie strength for the group of devices that are present in the location at a given moment in time. That is, we want to find methodology M(mt, st, t), in which *mt* is a mobility trace, *st* is a collection of pairwise social tie strengths for devices in *mt* and *t* is a time stamp in the range of time stamps of *mt*, and whose output is an aggregate social tie strength score.

1.3 Research Question

Creating the proposed methodology involves answering the following research question:

Is it possible to use patterns in device mobility data to infer the social connectedness for a given location and moment in time?

Answering this question will involve attempting to create a method that determines the strength of a social relationship between users, based on only location data, timestamps, and device identifiers. After these strengths have been determined, the social connectedness can be determined based on the social ties between all visitors in the location at the given time. This mobility method will be trained and tested by using a dataset of Wi-Fi access probe messages. The main benefit of using only mobility data is that it is available to a wide range of technologies and protocols, making the method very generally applicable. For example, the method could not only be applied to Wi-Fi data, but also to GPS, Bluetooth and video camera data. The proposed method would enable the automatic discovery of nearby social networks, which would benefit several applications, such as providing a statistical score for social behavior, improving MANET routing algorithms [31, 21] and enabling social user applications. One specific application of this method would be to support the Living Smart Campus project, whose goals include enabling crowd monitoring, while taking the privacy of users into consideration [47].

Our hypothesis is that this is possible to a significant degree, as one would expect that people with stronger social relationships are more likely to visit the same location at the same time. This hypothesis is also supported by the well-supported sociological theory of homophily [33], which states that socially connected people tend to be similar. This similarity could, for example, express itself as a similarity in mobility routines (e.g., colleagues going out for lunch at the same time) or as an interest in the same events (e.g., friends going to the same cultural performance).

The main research question will be answered based on the following sub research questions:

The first research question is: *Which mobility data features are correlated with social tie strength?* Previous research has used the similarity between the lists of previously accessed Wi-Fi networks as an indication of a link between devices [15]. These studies have used various metrics based on this similarity; this research question will be answered by selecting the metric that is correlated with social tie strength the most.

The second research question is: *Is it possible to accurately predict these social tie strength metrics from general (not Wi-Fi specific) device mobility data?* While similar lists of accessed networks indicates a link between devices, these lists cannot be obtained from general device mobility data. Many communication protocols other than Wi-Fi do not broadcast this information, and even Wi-Fi enabled devices may not always do so. This research question will be answered by constructing a method that predicts the similarity metric from the first question based on general mobility features, and then computes the aggregate social connectedness for a given location and time.

1.4 Main Challenges

The proposed method will be trained and evaluated based on a data set containing mobility information for devices that are identified by their (anonymized) MAC addresses. One difficulty related to this is the existence of randomized MAC addresses. Newer operating systems such as iOS 9, Android 6.0 and Windows 10 can prevent the tracking of user devices by periodically randomizing their MAC addresses. This causes multiple addresses to correspond to the same device, making it impossible to determine the actual mobility trace for these devices. This problem will have to be solved in order to create a useful method.

Secondly, it is not immediately clear how devices or people can be linked based on mobility information alone. People with a social connection will not always visit the same location for the same period of time and people may visit the same location independently, without having any social connection. Also, overlapping visits may be only weakly correlated with social ties, as some devices will overlap with many others simply because they stay in the location for a long time (e.g. those of staff members). Solving this problem will require the creation of a new method that uses multiple aspects of information about user mobility, in order to improve the accuracy of the prediction as much as possible.

Finally, the social network and aggregate connectedness computed by the method need to be validated by ground truth data. While validation by surveying users about their social relations would be preferred, this is impractical given the large number of users and resource constraints. Because of this, validation will need to be done based on the same data set as the one used to inform the mobility method.

2

Background

In this section, we describe the technology and techniques that the proposed method is based on. We also describe previously performed studies into related techniques. Specifically, we describe Wi-Fi and its usefulness for identifying devices and different metrics of similarity between lists of identifiers.

2.1 Wi-Fi

The proposed methodology is evaluated based on a data set of Wi-Fi access probes of visitor's devices. The data set is used to both generate the mobility trace, and to determine the social tie strengths between individuals. Here, we describe how the protocol works, and which information can be derived from these probes.

Mobile devices can connect to other devices by a variety of protocols, ranging from protocols for short distance communication such as Bluetooth, to protocols for longer distances such as LTE Advanced. Since 2015, the most popular type of communication by monthly offload traffic is Wi-Fi [13], which is based on the IEEE 802.11 standards [22].

2.1.1 Wi-Fi protocol

By using the 802.11 protocol, mobile and stationary devices can form a wireless local area network (WLAN), allowing access to the internet (in infrastructure mode) or inter-device communication (ad hoc mode). In infrastructure mode, mobile clients (known as Mobile Stations) communicate directly via radio with access points (APs), forming a Basic Service Set (BSS). By connecting multiple APs through a wired network multiple BSSs can be extended to an Extended Service Set (ESS), as shown in figure 2.1. When a mobile device leaves one AP and subsequently enters another, a *Handoff* is performed [35]. This mechanism consists of two processes: *Discovery*, in which the mobile device searches for nearby access points, and *Reauthentication*, in which the device and a selected AP exchange information and the device enters the new BSS.

Communication between 802.11 enabled devices occurs by using datagrams called *frames*, which consists of a number of MAC header fields (Frame Control), the payload (Frame Body), and a frame check sequence (FCS) [9]. The structure of a frame is displayed in figure 2.2. Frame Control consists of a number of smaller fields, among which the frame type and subtype, which together determine the category of the frame (e.g., beacon frame or probe request frame). Each address field contains a MAC address, identifying the devices on the path between the receiver and the transmitter.

Each MAC address consists of 48 bits, commonly represented by 6 octets [23]. The least-significant-bit of the first octet indicates whether the frame should be received by one or multiple devices. The second-least-significant-bit of the first octet signi-

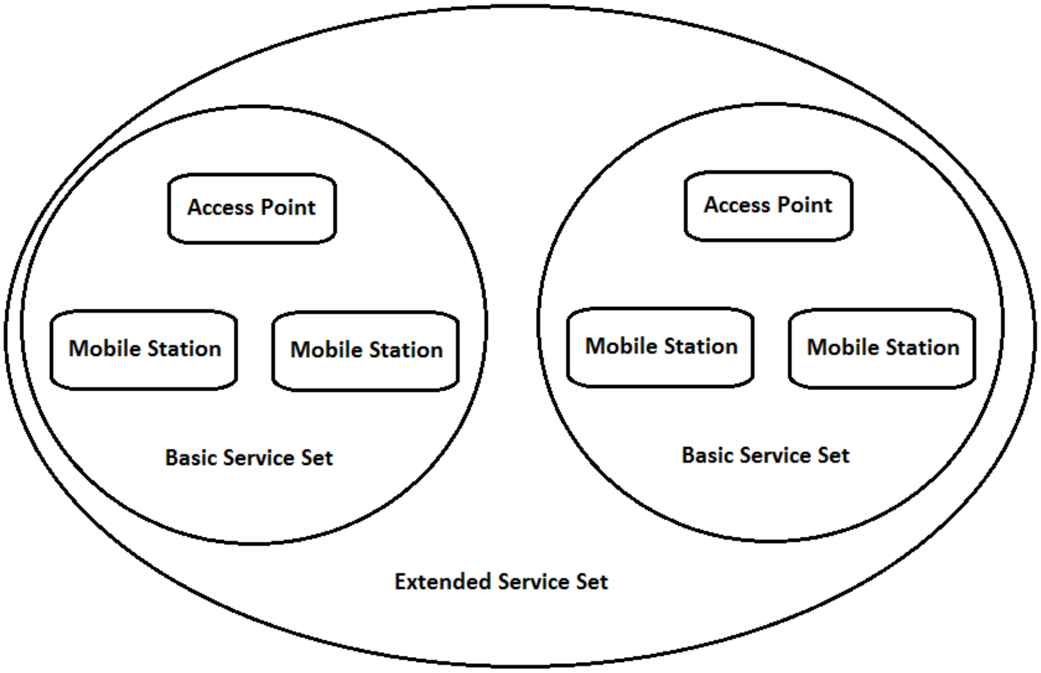

Figure 2.1: Organization of an Extended Service Set

fies whether the address is *globally unique* or *locally administered*. If it is globally unique, the address was assigned to the device by its manufacturer, and the first three octets identify the device manufacturer. These manufacturer identifiers are known as *Organizationally Unique Identifiers* or OUIs, and are administered by the IEEE Registration Authority. Otherwise, the address is locally administered, meaning that it has possibly been set by software on the device, and that the first three octets do not identify any organization.

Frame Control	Duration	Address 1	Address 2	Address 3	Sequence Control	Address 4	Frame Body	FCS

Figure 2.2: Structure of an 802.11 frame

During Discovery, the mobile client has two options to find a suitable access point. The first is *passive scanning*, which involves the device listening for so-called *bea-*

con frames that access points send out approximately every 100 milliseconds. The device can determine the signal strength of the received frames, and use this information to determine its preferred access point. The beacon frame contains an identifier known as the Service Set Identifier (SSID), which is often human-readable. Unlike globally unique MAC addresses, SSIDs can be set to an arbitrary string by the network manager, and often have some human significance. SSIDs often signify the manufacturer of the access point, the organization providing internet access through the access point, the organization that owns that access point, or some message that the network manager wants to transmit to passers-by. Common examples of SSIDS in the Netherlands are: ”linksys”, ”eduroam”, ”Ziggo” + a unique identifier and ”WiFi in de trein”.

The second method is *active scanning*. When using this method, the client device broadcast a message known as a *Probe Request* on a channel, and waits until an access point responds with a *Probe Response* [29]. The main benefit of this method is that it increases the speed and lowers the energy consumption with which Handoff takes place, as the client does not need to wait until the AP broadcast its beacon frames. Because of this, active scanning may be necessary for low latency applications such as VoIP. In practice, many devices even use active scanning when it is unnecessary, because the Wi-Fi implementation is often unaware of the requirements of the applications on the client device. According to a previous study [6], most Android and iOS devices perform active scanning about every 130 seconds. We obtained similar results by analyzing the activity of two devices running Android 6.0: both devices performed active scanning every 120 seconds with their screen turned off, and approximately every 10 seconds when their screen was turned on.

Request probes contain the SSID of the AP to which the client wishes to connect. If the client does not wish to connect to any particular AP, it can leave the SSID empty, in which case all nearby APs can respond. This is known as a *broadcast active probe*. The alternative is to specify a single SSID, known as a *directed active probe*. A device can probe multiple APs by simply sending multiple directed probes. Beside the increased Handoff speed, the main application of directed probes is for accessing networks that do not broadcast their SSIDs. Network managers sometimes feel that broadcasting SSIDs impacts their security or privacy, so access points often have the option to broadcast beacon frames without SSIDs, a practice which is sometimes called *network cloaking*. The AP will still respond when using directed active probes, so this method does not prevent usage of the access point altogether.

The SSIDs that a device uses during directed active scanning are selected by the operating system. Most operating systems store commonly used SSIDs in a list known as the *Preferred Network List* (PNL) [42, 2], or sometimes as the *Configured Network List* (CNL) [14]. This list is used by Windows, Mac OS, GNU/Linux, and mobile operating systems. As explained in the next section, the set of SSIDs from this list can often by used to identify the client device.

2.1.2 Device and Owner Identification

Existing studies show that information present in Wi-Fi packets can be used to identify both devices and their owners. [14] links devices to their owners by physically stalking them, while [7] shows which personal information can be derived from (mainly) the access points identifiers broadcasted by Wi-Fi devices (Preferred Network Lists or PNLs). Both articles focus on the privacy impact of these methods, and the second article offers some practical suggestions to counter this. [11] offers an alternative perspective on these methods: they can also be used to inform forensic investigations, by determining characteristics from the owners of devices present at the scene.

Nowadays, many operating systems allow users to enable MAC address randomization [48], which prevents tracking of the device by examining only MAC addresses. Unfortunately, each operating systems has a different implementation of randomization, which makes it difficult to determine how it impacts the MAC addresses in a real-world dataset. Implementations differ in terms of the requirements for using randomization, when it is used, and how the address is randomized. Android, Windows and Linux require that the hardware and drivers support randomization. Most operating systems only use randomization when scanning, while Windows can also use a random address when connecting to an access point. Both Windows and iOS [36] set the *locally administered* bit when using randomization, but we could not find any literature describing the same behavior for Android and Linux.

2.2 Similarity Metrics

In order to train our method to infer social tie strength from mobility features, we must first know what the actual tie strength is for each sample pair of devices. One way to do so would be to ask each device owner how well they know each other owner, but this would be impractical for any group of significant size. However, previous studies have shown that tie strength can also be inferred based on information present in Wi-Fi request probes, which is much more scalable [15]. Specifically, the methods in these studies gather the SSIDs that were broadcasted by both devices (PNLs) and compute a metric indicating the similarity between the two resulting lists. The idea of this method is that the more similar these SSID lists are, the greater the overlap of previously visited locations is, and the more likely that the device's owners have a social relationship is.

In order to select a suitable metric, we examined different metrics that have been proposed in the literature. Many of the metrics were originally created in the field of information retrieval, where they were used to assess the similarity between one sequence of words (e.g., a search query) and another (e.g., a web page). As such, some of the metrics consider the *term frequency* of a word, which is equivalent to the number of times that the word occurs in a particular sentence. Since SSID lists are sets, each SSID occurs only a single term, so it does not make sense to consider term frequency. Therefore, we also look at modified variants of one particular metric (TF-IDF) that do not use term frequency. A second measure that some of the metrics use, is the *document frequency* of a word, which is the number of documents in the whole collection of documents (the corpus) that the word occurs in. In our comparison, this refers to the number of times the SSID occurs across all SSID sets.

We will examine the following metrics: the word overlap fraction, the Jaccard Index, TF-IDF and a number of its variants, and Adamic-Adar and one of its variants. After doing so, we will explain which of these metrics we chose to use to infer social tie strength from PNLs.

2.2.1 Word Overlap Fraction

The simplest metric of sentence similarity is word overlap fraction [34], which is the proportion the words in a query that are present in the considered sentence: $S(Q, R) = \frac{|Q \cap R|}{|Q|}$, where S and Q are sequences of words or identifiers. Since this metric considers one sequence (the query) differently than the other (the considered sentence), it is not suitable for our application.

2.2.2 Jaccard Index

A similar metric that can be used is the Jaccard Index [15], which is the proportion of words in either sequence that are present in both sentences: $S(Q, R) = \frac{|Q \cap R|}{|Q \cup R|}$. The function's range is [0, 1]. One possible problem with this metric is that each overlapping identifier contributes equally to the resulting value, regardless of its rarity. Also, the number of overlapping identifiers does not necessarily increase its value, as both sets may be small. As an example, consider the sets A = {eduroam} and B = {Ziggo914781, VGV8128421}. Using the Jaccard Index, S(A,A) = S(B,B) = 1, because in both cases each list contains all identifiers present in the other. However, S(A,A) should intuitively result in a much lower score than S(B,B), because A contains a single common identifier, while B contains multiple very uncommon ones. Therefore, the Jaccard Index is also unsuitable for our application.

2.2.3 TF-IDF

Now we consider a number of metrics that do take into account word rarity and the number of overlapping words. The first is TF-IDF, short for term frequency - inverse document frequency, which was first presented in [45]. It has many formulations. We use the one present in [34]: $S(Q,R) = \sum_{w \in Q \cap R} log(tf_{w,Q} + 1) log(tf_{w,R} + 1) log(\frac{N+1}{df_w + 0.5})$, where $tf_{w,Q}$ is the term frequency of word w in sentence Q, df_w is w's document frequency, and N is the total number of documents (in our application, the number of SSID lists). The function's value has a lower bound of 0, but has no upper bound, as the number of overlapping terms and their frequencies can be arbitrarily high. The intuition behind the function is that similarity should be increased the more the overlapping terms occur in either sentence (the higher the term frequency), and the rarer that the terms are across the whole corpus (the inverse document frequency). However, since term frequency is not useful for our application, the function can be simplified to the following: $S(Q,R) = \sum_{w \in Q \cap R} log(\frac{N+1}{df_w + 0.5})$.

An additional TF-IDF variant that is modified to be used for set similarity is presented in [15], which is computed as the cosine similarity between the vectors of the inverse document frequency of the words in the two sets. The measure, called Cosine-IDF, is computed as follows: $S(Q,R) = \frac{\sum_{w \in Q \cap R} IDF_w^2}{\sqrt{\sum_{w \in Q} IDF_w^2}\sqrt{\sum_{w \in R} IDF_w^2}}$ with $IDF_w = log(\frac{1}{df_w})$. Its value ranges from 0 to 1. The metric suffers from the same problem as the Jaccard Index: as long as the two compared sets have the same members, the resulting score will be 1, regardless of their member's rarities.

2.2.4 Adamic-Adar

Another metric is Adamic-Adar similarity, which was originally used in [1] to infer social relationships between users from the similarity of their personal web pages. It is computed as the summation of the inverse document frequencies of overlapping terms: $S(Q,R) = \sum_{w \in Q \cap R} \frac{1}{log(df_w)}$ A variant of this metric is presented in [15] which is called Psim-q: $S(Q,R) = \sum_{w \in Q \cap R} \frac{1}{df_w^q}$. Here, q is an extra parameter, that determines the effect that rarer overlapping terms have on the similarity score. The referenced study evaluated the metric for multiple values of q, and found that it works best when equal to 3. Like TF-IDF, both have a lower bound of 0 and no upper bound. Unlike Cosine-IDF, the score for two sets with the same members is higher if the member SSIDs are rarer.

2.2.5 Conclusion

Many of the discussed metrics have been used in recent studies in order to infer the strength of social links from SSIDS: for example, [4] uses Adamic-Adar, [10] uses Cosine-IDF, and [30] uses TF-IDF. In order to select the one best suitable to our application, we considered the results of a previous study [15], which evaluated the performance of multiple metrics when trying to predict known social links. In the study, both the Cosine-IDF and Psim-3 metrics had a high accuracy. Because Psim-q does not suffer from the same problem as Cosine-IDF with regard to taking into account the rarity of SSIDs, we chose to use Psim-3 as our similarity metric.

3
Related Work

In this section, we review existing literature on subjects related to the purpose of this study. Specifically, we look at previous studies on the subjects of mobility modeling and social links between devices.

3.1 Mobility Modeling

There have been many previous studies dealing with Wi-Fi based and more general location tracking and prediction. One of the earliest studies focused on Wi-Fi location prediction is [26], whose method predicts only aggregate movements. Later studies such as [49] and [44] outline several methods that have broader utility, in that they predict the next access point that devices will visit based on their previous movements. [37] describes a method that is even more elaborate, which gives a probabilistic prediction of the geographical device position, and not just the access points. We have also examined papers that describe more general location prediction methods, that are not linked to the Wi-Fi protocol, but are none the less usable for Wi-Fi location prediction. The first of these, [50], only uses previous locations to inform the prediction algorithm, but the following papers use additional information such as timestamps ([18]) and social links ([32]). Finally, [17] presents a general prediction method that can use any combination of input features.

The paper [26] aims to answer the research question: ”Can clustered hourly Wi-Fi activity be used to model aggregate user movements between access points?” The study aims to create a model of user mobility, based on real-world data. This research was done in a time that laptops, not smart phones, made up the majority of Wi-Fi enabled devices. Since laptops are mostly used when stationary, this made it impossible to model actual device trajectories, which is why the paper focuses on aggregate influx and outflux at access points. The model is created based on a dataset of Wi-Fi packets from almost 14000 devices for 2 months. After aggregating the records by hour of the day, they are divided into 5 clusters with similar activity patterns. Finally, for each of the clusters, the daily arrival and departure rates are computed and synthetic traces are generated. A benefit of the work is that it modelled aggregate user movement between access points, even though individual user movement was unknown. The main difficulty of the work was to cluster access points with similar activity patterns and to generate synthetic traces. One major limitation is that the model is not evaluated, which the paper mentions as future work.

In [49], a method for predicting the movement of people is described, which is implemented by the Jyotish framework. Its research question is: ”Based on combined Wi-Fi and Bluetooth data, is it possible to predict where a person will stay, for how long, and who they will meet?” The method works by having user devices collect Wi-Fi records (indicating location) and Bluetooth records (indicating user contact). The Wi-Fi records are then used to determine the location for each Bluetooth record, and the combined Bluetooth records and locations are then used to construct a predictor for each of the 3 sub research questions. The method was evaluated by 50 users over 20 to 50 days, and had a high accuracy for each of the predictors. The main benefit of the method is that it uses multiple sources of data in order to predict both user location and contact. The main difficulty of the research was in determining the user location based on observed access points and then assigning these locations to the Bluetooth data. Additional difficulty was involved in creating the 3 predictors from the combined data. The main limitation is that the method requires that individuals use have Bluetooth enabled in order to determine user contact, while this may often not be true in practice.

The paper [44] describes an empirical comparison of various location predictors that were previously described in the literature. Its research question is: ”How do Markov-based predictors perform in comparison with compression-based predictors when predicting future user device locations based on sequences of used Wi-Fi access points?”. The study considers algorithms from two families of domain-independent location prediction families: Markov-based and compression-based predictors. These algorithms are compared on the basis of their accuracy of predicting the next location given a trace of previous locations. The study uses a dataset generated by 6000 users over 2 years to perform the evaluation. It found that the O(2) Markov predictor had the highest accuracy. The main benefit of the study was that it applied existing algorithms to large-scale real world data, which had not been done before. The main difficulty was that each predictor needed to be adapted so that they would perform well on the Wi-Fi mobility data. Limitations include the fact that only a small number of algorithms were tested, and that they were compared based on only their accuracy.

The authors of [37] describe a number of methods to track Wi-Fi enabled devices and to estimate their geographical trajectories, instead of just determining which access points they have used. Its research question is ”Can the trajectory of devices along roads be estimated from sequences of previously used Wi-Fi access points?” The main contribution of the paper is a probabilistic Hidden Markov Model-based method that estimates the trajectory of devices based on possibly sparse Wi-Fi transmissions at possibly sparsely or densely distributed access points. It also describes multiple methods to increase the number of Wi-Fi transmissions per device, which increase the number of detected devices and the location accuracy for each device. These methods are then evaluated by calculating the difference between the estimated and actual device locations (based on GPS) under various configurations. The results show that the method has high accuracy, which degrades gracefully as the density of access points is decreased. The main benefit of this estimation method is that it seems to perform well under less than perfect circumstances.

The study [50] analyzes a method for cell-based location prediction, which applies to both GSM and certain configurations of Wi-Fi access points. Its research question is: ”Can a data mining algorithm accurately predict inter-cell user movements from previous user paths?” The proposed algorithm has three steps. In step 1, a data mining algorithm extracts patterns from previous sequences of user inter-cell movement, where each pattern has the form <c1, c2, ..., ck>. In step 2, these patterns are converted to mobility rules of the form <c1, c2, ...> → <..., ck>, together with their confidence values. In the final step, the mobility rules are applied based on the current path, producing a list of predicted paths sorted by confidence plus support of their generating rules. In the evaluation, the precision and recall of the algorithm is compared to those for two other methods: the Mobility Prediction based on Transition Matrix (TM) method and the Ignorant Prediction method. The proposed method has higher precision than the other two methods, while it has a lower recall for most prediction counts. The largest benefit of this method is therefore that it has a higher precision than comparative methods. The first step likely involved the greatest difficulty, and the other steps seem fairly straightforward. A limitation to the study is that it only considers previously visited locations as input variable, and does not consider other contextual information.

Similarly to the previous study, [18] attempts to create a data mining method to predict cell-based movement. The main difference is that it also takes the time of day in account to create and apply rules. Its research question is: ”Can a data mining algorithm accurately predict inter-cell user movements from previous user paths and times of day?” Instead of representing movement as sequences of cells, this method represents movement as sequences of (timestamp, cell) tuples. As in the previous study, patterns are mined from previous user movements, which are converted to rules with confidence values, with the form <(c1, t1), (c2, t2)> → <(c3, t3), ..., (ck, tk)>. The confidence values also take into account how long ago the movement was performed: more recent sequences produce rules with higher confidence values. The described method is evaluated by applying it to a synthetic dataset and by varying multiple algorithm and dataset parameters. The results show that the method has a high precision and recall under the majority of circumstances. The main benefit of this study is that the method uses time of day to improve its predictions, which seems to have a positive effect. One limitation of the study is that the method's results are not compared to those of other methods, so it does not effectively demonstrate that the method is better than existing ones. A second limitation is that the method only uses time of day to inform its rules, while other time-based features are ignored, such as day of the week and season.

The paper [32] describes a method that incorporates the strength of social links to improve location extraction. Its research question is: ”How can different features of social relationships be used to improve location prediction of social network users?” The research aims to improve upon a previously reported method that used social links to predict users' home locations, but did not take into account the strength of those links. The study is also based on a dataset from the Twitter social network; because its users tend to follow accounts that are not close geographically (such as celebrities), this is a major issue. The study starts by examining different factors of social links to see which correlate with distance between contacts, with the following main results: reciprocal friendships indicate closeness more so than weaker types of links, users tend to be closer to users with private accounts, and users tend to be further from accounts with many followers. After this, the study creates a decision tree regressor to use these features to determine which users are likely to be close. An evaluation of the resulting predictor shows that the method has a higher accuracy than the existing method. The main benefit of the study is that its method successfully uses social tie strength to improve location prediction accuracy. A limitation is that the selected features and the predictor's results are strongly dependent on the used dataset, and would likely be significantly different for social networks other than Twitter.

Whereas the previous studies only considered specific parameters (such as location and time of day) as predictive parameters, [17] describes a general method to predict location based on any number of contextual parameters. Its research question is: ”Can the combination of multiple contextual models be used to accurately predict the prediction of user location and visit duration?”. In the proposed method, values for each input variable are generated by separate ’contextual models’, and the outputs of these models are combined in order to predict the output variables, which in this case are location and visit duration. In addition to location and time, contextual variables could include (for example) the application logs on the user's smartphone, and the density of nearby Bluetooth devices. The outputs of the contextual models are combined by an ensemble method, in which multiple combinations of model outcomes are weighted and then multiplied to compute the output values. The method learns the weights for each individual based on a training dataset. The study evaluated the method by creating methods for predicting location and visit duration. Contextual variables for both tasks included the current location, hour of the day, day of the week, whether the day is a workday or weekday, and frequency of visits to the current location, and other features. The results of the evaluation showed that both methods have a high prediction accuracy, and that the location prediction model improves as the number of location transitions increases. A large benefit of this method is that it is general enough to capture existing methods (like those from the previous two studies), and allows for the inclusion of previously unconsidered variables.

3.2 Social Link Prediction

The papers in this section describe methods that infer social links between users from either Wi-Fi specific information, or more general features (such as social proximity and visited locations). [15] looks at multiple methods to determine the similarity between SSID lists, which often indicates a relation between users. [10] also looks at multiple techniques to infer a social link from Wi-Fi data, including similar SSID lists, physical proximity, and overlapping visits. [4] uses SSID list similarity to extract a social network, and uses it to confirm the sociological theory of homophily. This is the theory that people with social connections tend to be similar [33], and it is supported by many studies. By applying homophily, [30] improves venue recommendations based on venues visited by socially linked users. One other study [16] looks at aggregate user behavior: the presented method extracts the home locations of visitors and uses this data to accurately predict election results. Finally, the methods in other studies infer social links not from Wi-Fi data, but from the social network in the past [28] and the visit distribution to commonly visited locations [41].

The study described by [15] examines a method that aims to determine social relationships between users from the SSID lists that their Wi-Fi devices broadcast, with the research question: ”Can social links between device owners be inferred from overlapping lists of preferred SSIDs?” In order to determine the best method to determine the existence of social links from SSID lists, the study implements 4 similarity metrics and compares their performances. These metrics are then tested by using a dataset from 8000 devices. The study finds that a cosine-IDF metric and a modified Adamic-Adar similarity metric have the best performance in terms of true and false positives. Finally, the paper suggests several countermeasures for the possible privacy impacts of this method. The main benefit of the study is that it demonstrates the effectiveness of the two metrics by using a large dataset. The main difficulty was in finding and implementing suitable metrics, and in analyzing which metrics performed the best.

The paper [10] describes a study that attempted to infer social relations between users based on the activity of their Wi-Fi enabled devices. Its aim is to answer the research question: ”Can social relationships be inferred from Wi-Fi data indicating previously used networks, physical proximity and spatio-temporal behavior”? As indicated by the research question, the study considers 3 separate techniques. In the first technique, users are considered similar when their devices have similar lists of previously used access points (PNLs), which is computed by a Cosine metric. In the second technique, user PNLs are converted to a list of previously visited locations by mapping SSIDs to geographical coordinates (based on wardriving databases). If users share at least one location, the first technique is applied. The third technique determines the probability that users are in the same location at the same time, by using a local monitoring system. Each of the techniques has been successfully demonstrated by experiments. The benefit of this study is that it provides and confirms the use of several promising techniques to infer user relationships.

The research presented in [4] aims to create a method in order to answer the research question: ”Can a social network be determined from Wi-Fi request probes, and based on this network, can the strength of social relationships be linked to usage of the same device types?” The described method links users based on their similar preferred network lists (PNLs), by using the Adamic-Adar metric. After doing so, it creates a social network, whose properties are analyzed by the paper. The method is evaluated by using multiple datasets with Wi-Fi probes from more than 9000 devices each. The paper finds that users with social relationships are significantly more likely to use devices from the same vendor and to use the same language. The main benefit of the study is that it demonstrates how a social network can be determined from PNLs, and that it confirms the sociological theory of homophily (physical proximity is related to interconnected traits). The main difficulty of the study was in selecting the Adamic-Adar metric among several metrics, and to analyze the various properties of the social network.

The next paper, [30], describes a framework and method for calculating a personalized venue reputation score for users that takes into account the activity of other users that have a similar list of preferred Wi-Fi access points (PNLs). The study hopes to answer the research questions: ”Can individual reputation scores for physical venues be determined based on PNL similarity with other users?” In the proposed architecture, Wi-Fi access points for a certain venue collect the PNL and visitation frequency for each visitor. It then calculates a TF-IDF based similarity score between PNLs in order to determine how close visitors are socially. By combining the similarity scores and visitation frequencies, the framework determines a score that indicates how likely the visitor is to be interested in the venue, which is transmitted to the user device. Since this is a position paper, it does not describe the whole study (which will be completed in the future). At this point, the main difficulty was in finding measures for user similarity and expected interest. A benefit to this approach is that it would automatically take into account the interest from a large subset of venue visitors, while review systems can only take into account the interest from a small subset of visitors. A second benefit is that it uses the social relation between users to improve the reputation score, which many existing systems do not.

The study presented in [16] attempts to use Wi-Fi data to extract social information from crowds, with the research question: ”Can Wi-Fi probe request data (specifically PNL SSIDs) be used to determine the geographical origin and associated information for large groups of users?” By comparing PNL SSIDs to wardriving databases, the study infers the likely location that a user came from. This method was applied to datasets from events with varying degrees of geographical distributions of their visitors: international, national, and city-wide events. The study analyzes the geographical provenances from these datasets, and infers several reasons for their distributions, based on the location and types of events. Finally, the study combines the geographical origins from a nation-wide political event with city-based election data, in order to attempt to predict the election outcomes at the event. The results of these predictions are highly accurate, suggesting that Wi-Fi probe data can be used to infer the political leaning of crowds. The main difficulty of this research was in mapping each list of SSIDs to the most likely city of origin, which required the use of a provenance rank based on wardriving datasets. While this method has been applied in previous studies, the main contribution of this research is its successful prediction of election data, which may have larger societal implications.

The study in [28] compares the performance of various methods on the social link prediction problem. This study aims to answer the question: ”How do existing algorithms perform when attempting to predict future social links from the current social network?” The general idea behind these predictors is that individuals that are close in a social network (social proximity) are more likely to form social link in the future. The research evaluates the methods by applying them to five datasets from coauthorship networks, from different moments in time. Since the accuracy of each method is low, their performance is compared to that of a random predictor. The results show that the Adamic-Adar algorithm has the highest average accuracy, although some other methods (Katz clustering, common neighbors) have a very similar performance. The study also finds that the performance of all methods (relative to random) improves when applied to larger social networks. The main benefit of the research is that it gives an overview of the performance of multiple algorithms, and analyzes why they perform similarly. A limitation is that the evaluation only describes the accuracy of the methods, and does not compare other performance metrics, such as precision and recall.

Finally, [41] describes a study into a technique to improve social link prediction, focusing on users of location-based social networks, in which users ’check-in’ to locations they visit. Its research question is: ”How do we design a link prediction system which exploits data about user check-ins?” The paper describes a framework that uses supervised learning to predict social links based on a number of features. This includes two location-based features for each pair of users that have visited the same location: the minimum place entropy across all venues they have both visited, and the sum of the inverse of each place entropy value. Place entropy is a metric that indicates how evenly distributed the number of check-ins per user is for some venue; a lower place entropy indicates that a small number of users has a high number of check-ins. The algorithm also only considers users that have visited locations or friends in common, considerably reducing the search space. The paper evaluates the performance of multiple classifiers on a dataset from the Gowalla social network, and finds that model trees and random forests have the highest AUC, precision and recall. The main contribution of this study is that it shows that location features can improve social link prediction. One limitation is that it only considers two features (both based on place entropy), and does not consider others such as venue category and opening and closing times.

3.3 Conclusion

As reviewed in the previous section, multiple previous studies have successfully extracted social networks from Wi-Fi data, and have used this network to infer other types of information. However, the main problem of these methods is that they can only be applied to Wi-Fi data, and not to general mobility data such as GPS, cellular and Bluetooth data. This is the main problem that our proposed method could contribute to solving, as it is based on only mobility features. One exception to this problem is one of the methods presented in [10], which describes a mobility feature that is generally applicable. However, this feature misses many sources of information, and thus is unlikely to be a good predictor on its own. In this study, we will create a method that takes into account a large variety of mobility features, which we suggest will result in a method with more accurate predictions.

4
Research Method

In this chapter, we describe the workings of both a state of the art baseline methodology and the proposed methodology, whose performances we will compare in the results section. We first describe the baseline methodology, which is based on a mobility feature that is described in a recent study. We also describe the main problem with this method, and why we chose to develop an alternative method. Then, we describe the different phases of the proposed methodology. The method's first phase uses a feature selection algorithm to select the most promising mobility features among a larger number, and uses them to train a model. The second phase uses this model to predict the aggregate social tie strengths for a location, which is the methodology's output.

4.1 Baseline Methodology

In order to predict social tie strengths from mobility data, we have to first mobility features that are correlated with social connections. One promising feature is based on one of the methods used in a previous study [10], where it is called ”spatio-temporal co-occurrence probability”. The feature defined as the probability that both users are in the same location at the same time. The study uses this feature based on the belief that visitors with a social relation are more likely to meet each other than unrelated visitors.

For a single location, the feature can be defined as follows. Given a vector V_i for each device i, in which each entry V_{it} is equal to 1 if the device was present in the location at time slot t, and 0 otherwise, then the feature is defined as: $\frac{\sum_t V_{1t} * V_{2t}}{\sum_t V_{1t} * \sum_t V_{2t}}$.

For example, given $V_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}$ and $V_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}$, then the feature is equal to $\frac{2}{5}$.

This is a promising feature, as it makes sense intuitively that people with a stronger social connection more often visit the same location and vice-versa. However, this feature misses many other sources of information that may inform the social tie strength prediction, for which other features may be defined. Example features are the total amount of times spent in the location by either device (which could indicate something about the behavior of the device), and the number of people present in the location during overlapping visits (if there are more people, visiting devices are less likely to be related). For this reason, we chose to develop an alternative method that uses multiple features of the mobility trace in order to infer social links.

In order to evaluate the performance of the proposed methodology, we will compare it to a baseline methodology based on this feature. This baseline feature is also used as one of the features of the proposed methodology. Since the feature is related to both the number of overlapping visits and the overall number of visits of both devices, it is part of the ”Overlap and Individual” feature class. Because the proposed method can use any combination of the proposed features, including only the baseline feature, it should always perform at least as well as the baseline method.

4.2 Proposed Methodology

The general approach of our method is as follows: there are two phases, during which the model is learned and then applied. The first phase is the initialization phase, in which the model is trained and its features are selected based on a mobility trace and knowledge about the pairwise social tie strengths (which we infer from the similarity of Wi-Fi SSID lists). This is followed by the utilization phase, in which the mobility trace is supplied to the model, generating predicted pairwise tie strengths. By combining these strengths with the devices present at a given time stamp, an aggregate social connectedness score is then calculated and outputted.

The diagram in figure 4.1 gives a simple overview of the proposed method.

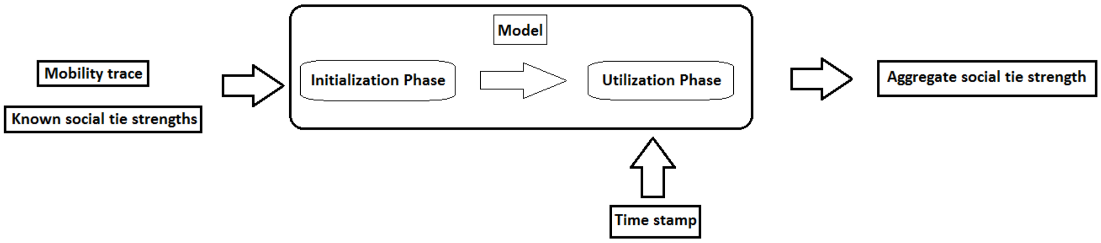

Figure 4.1: High-level steps of the proposed methodology

4.2.1 Initialization Phase

To create our model, we used supervised machine learning to train a regressor from samples consisting of mobility features as input features and with social tie strengths as labels. Each sample, which represents a pair of devices, has the following form:

$(tie\_strength, feature\_value\_1, feature\_value\_2, \ldots, feature\_value\_n)$, where n is the number of used features.

Both the tie strength and the mobility features are computed based on a data set of Wi-Fi access probes, in which each device is identified by their MAC address. However, MAC addresses can be randomized, so a single device can actually correspond to multiple addresses, which is problematic. One way to deal with this issue is to attempt to defeat this randomization and link the addresses to a single device, which previous research has shown to be possible [48]. However, we decided not to do so, as this would likely go against the wishes of privacy conscious users that have enabled this randomization. Instead, we tried to detect randomized addresses and remove them from the data set. We did so by examining the second to last bit of the first octet of the MAC address, which indicates whether the address is locally administered or not, and many implementations of randomization set this bit [36]. One problem with this approach is that the accuracy of the predictions could decrease by removing devices from the data set. However, we found that in practice only a small number of devices employ randomization (as shown in the results section), reducing its impact.

The tie strength for each sample is determined by comparing the SSID lists for the pair of devices, and by computing a value that measures their overlap. We use this method because its value has been shown by multiple previous studies [15, 10, 4, 30, 16], and because these values could be derived from the dataset available to us. There are multiple metrics that can be used to compute the similarity between SSID lists, which perform differently for the number of overlapping identifiers and their frequency in the data set. We specifically used a modified version of the Adamic-Adar metric known as Psim-3, because [15] showed that it outperformed most other metrics in determining links between individuals. It is calculated as $\sum_{z \in X \cap Y} \frac{1}{f_z}$, in which X and Y are the two SSID sets, and f_z is the number of times that identifier z occurs in the data set.

The mobility features are computed based on the time stamps of Wi-Fi request probes, which identify at which moments in time the device was near the location. However, many features are based on amount of time spent at the location, so the time stamps cannot be used directly. Instead, we convert each list of time stamps to a number of visit start and end time stamps. These visits are determined by grouping together time stamps that have at most a certain length of time between them (the maximum gap length). This gap length was chosen so that it was higher than 95% of the gaps (between 2 and 4 minutes in practice). We tested the method's performance for other ratios, but we found no consistent improvements.

As our regressor we used an implementation of a gradient boosting algorithm [51]. Specifically, we used the implementation by the Python Scikit-Learn package [39], version 0.18.1, with the default parameters and maximum tree length set to 8. We used this algorithm as it performed better than any other regressor in the package for many combinations of input features and training sets during our evaluation. Other regression algorithms implemented in the package include Linear Regression, Ada Boost, Bagging, Extra Trees, Random Forest and Multi-layer Perceptron. Boosting algorithms work by training a large number of simple models called weak learners and by combining their results. In each training round, the algorithm adds a new weak learner and reweighs the samples so that poorly predicted samples are better predicted by future weak learners. Gradient boosting differs from earlier algorithms such as Ada Boost mainly because they can be applied to any (differentiable) cost function. Its main benefits are that it performs well for complex hypotheses and that it is not very susceptible to overfitting.

Algorithm 1 shows the pseudo code for the first part of the initialization phase, in which the samples are generated. Its inputs are the mobility trace (a collection of detections with form <d, t>, with *d* being a device and *t* being a time stamp) and a collection of pairwise social tie strengths between some of the devices in the mobility trace. Its output is a collection of samples, each of which is a tuple <st, MF>, with *st* being a pairwise social tie strength for some pair of devices, and *MF* being the values of the mobility features for the same device pair.

Algorithm 1: Initialization Phase, Sample Generation

```
Data: <mobility trace MT, social tie strengths ST>
Result: collection of samples SC
SC = [];
D = determineDevices(MT);
DP = computePairs(D);
forall pair in DP do
    MF = computeMobilityFeatures(pair, MT);
    st = ST[pair]; /* the pairwise social tie strength */
    append(SC,<st, MF>);
return SC;
```

After computing the mobility features for each device pair, we determined which features should be supplied to the regressor. We decided to select these features by algorithm instead of using a fixed selection, because we found that the regressor had its optimal performance for different locations for different feature sets. The features are selected from a larger set of mobility features that we discuss in the next subsection. The feature selection algorithm can be classified according to the scheme in [27] as follows: its starting point is the empty set, it moves through the search space as a greedy algorithm, it evaluates features as a wrapper method (by determining the performance of the regressor), and it halts when no new features improve the regressor performance. The performance of the regressor was computed by performing 10-folded cross validation and determining the average of their mean squared errors. We chose to aim to minimize the mean squared error in social tie strength, because we value errors equally across the range of social tie strengths, as ties contribute equally to the aggregate value computed in the utilization phase. The result of this phase is the best performing regressor and its input features.

Algorithm 2 shows the pseudo code for the second part of the initialization phase, in which the feature selection takes place. Its input is the collection of samples generated in the first part of the initialization phase. Its output is a regressor trained using the combination of features as selected by the feature selection algorithm. Some important variables are: *FIC*, which stores the current best selection of feature indices, updated after each round; *BTFIC*, which stores the best selection of indices found during the round, and *s*, which stores the highest accuracy found among tested regressors.

Algorithm 2: Initialization Phase, Feature Selection

```
Data: collection of samples SC
Result: regressor r
FRC = range(0, length(SC[0])); /* range of all feature indices */
FIC = []; /* current best feature indices overall */
BTFIC = []; /* current best feature indices for the current round */
s = 0; /* current best score overall */
bts = 0; /* current best score for the current round */
sib = true; /* boolean indicating whether score has improved */
ftb = true; /* boolean indicating whether first score has been generated */
frb = true; /* boolean indicating whether first round has been completed */
while sib do
    forall index in FRC do
        if index in FIC then
            continue;
        TFIC = union(FIC, [index]); /* feature indices to test */
        TS = selectByIndices(SC, TFIC); /* samples to test */
        ts = crossValidateRegressor(TS, 10); /* score from test */
        if $ts > bts$ or $ftb$ then
            bts = ts;
            BTFIC = TFIC;
            ftb = false;
    if $bts > s$ or $frb$ then
        s = bts;
        FIC = BTFIC;
        frb = false;
    else
        sib = false;
r = trainRegressor(SC, FIC);
return r;
```

4.2.2 Features

We have selected 124 possible mobility features, which fall into a number of more general feature classes. Each feature class represents a type of information related to visits to the location, and each feature represents some specific measure related to their class. For example, for the class 'Overlap Only', features include the number of overlapping visits and the total length (in seconds) of their overlap. We chose these feature classes in order to group the features by their source of information (overlapping visits, devices themselves, and the environment). This will later allow us to reason about the importance of each source of information in predicting social tie strengths. The feature classes are as follows:

- Overlap Only These features relate directly to the overlapping visits for the pair of devices. Examples are the number and total length over overlapping visits, and the average length of time that one device waits before and after the other arrives. There are 51 features in this class.
- Individual Only These are based on visits the devices have made regardless of overlap. They include the total number and length of visits by the devices, and their average and median visit lengths. The class contains 16 features.
- Overlap and Individual These features relate some measure of overlapping visits to a measure of the devices' overall visit pattern. For example, it contains the ratio between overlapping visits and total number of visits, and the ratio between overlap length and total visit length. It contains 8 features.
- Overlap and Location The features in this class measure the state of the location when the devices had overlapping visits. Specifically, it considers how busy the location was at the time, in terms of the number of concurrent visitors compared to its maximum. Example features are the average and maximum popularity during overlapping visits. It contains 7 features.
- Individual and Location Similarly to the previous class, features in this one are based on the location popularity, but they consider all visits of both de-

vices, instead of only overlapping ones. Example features are the average and maximum popularity during visits. The class contains 42 features.

There is no class related to the location only, because its features would be independent of the pair of devices, and have the same values for each sample.

Feature Descriptions

Table 4.1 summarizes the specific features that each class contains.

The features are based on the mobility of two devices, which we will refer to as *d1* and *d2*. For each of these devices, we consider the set of visits to the considered location, *visits_d1* and *visits_d2*, which consist of tuples of the form <visit_start, visit_end>, in which both elements are a time stamp. Additionally, we consider the set of overlapping visits, *overlapping*, which is derived by determining which visits in *visits_d1* took place during a visit in *visits_d2*. It consists of tuples of the form <d1_arrive, d2_arrive, d1_leave, d2_leave, overlap_start, overlap_end>, in which each element is a time stamp. The elements *overlap_start* and *overlap_end* are derived from the other elements as follows: $overlap\_start = max(d1\_arrive, d2\_arrive)$ $overlap\_end = min(d1\_leave, d2\_leave)$. Finally, *overlapping_d1* and *overlapping_d2* are similar to *overlapping*, but each overlap is calculated from the point of view from a single device. For example, if a visit from device A starts during one visit of device B and ends during another, it counts as 1 overlapping visit for A and as 2 for B.

Additionally, we define some functions: $length(visits\_d1_i)$ for some *i* returns the number of seconds that the visit took (visit_end - visit_start). Similarly, $length(overlapping_i)$ returns the number of seconds of overlap (overlap_end - overlap_start). $range(t1, t2)$ returns the set of all time stamps from time stamp *t1*, up to time stamp *t2*. $present(t)$ returns the number of devices that are present in the location at time stamp *t*, and *max_present* is the highest count returned by *present* across the whole data set.

Table 4.1: Features by class

Feature class	Features
Overlap Only	$\lvert \mathit{overlapping} \rvert$, $\sum_i \mathit{length}(\mathit{overlapping}_i)$, $\{\mathit{overlap\_start} - \mathit{d1\_arrive} \mid \mathit{visit} \in \mathit{overlapping}\}$ [*] [†], $\{\mathit{d1\_leave} - \mathit{overlap\_end} \mid \mathit{visit} \in \mathit{overlapping}\}$ [*] [†], $\{\mathit{overlap\_start} - \mathit{min}(\mathit{d1\_arrive}, \mathit{d2\_arrive}) \mid \mathit{visit} \in \mathit{overlapping}\}$ [*], $\{\mathit{max}(\mathit{d1\_leave}, \mathit{d2\_leave}) - \mathit{overlap\_end} \mid \mathit{visit} \in \mathit{overlapping}\}$ [*], $\{(\mathit{overlap\_start} - \mathit{min}(\mathit{d1\_arrive}, \mathit{d2\_arrive})) + (\mathit{max}(\mathit{d1\_leave}, \mathit{d2\_leave}) - \mathit{overlap\_end}) \mid \mathit{visit} \in \mathit{overlapping}\}$ [*]
Individual Only	$\lvert \mathit{visits\_d1} \rvert$ [†], $\{\mathit{length}(\mathit{visit}) \mid \mathit{visit} \in \mathit{visits\_d1}\}$ [*] [†]
Overlap and Individual	$\lvert \mathit{overlapping\_d1} \rvert$ [†], $\lvert \mathit{overlapping} \rvert / \mathit{max}(\lvert \mathit{visits\_d1} \rvert, \lvert \mathit{visits\_d2} \rvert)$, $\lvert \mathit{overlapping} \rvert / \lvert \mathit{visits\_d1} \rvert$ [†], $(\sum_i \mathit{length}(\mathit{overlapping}_i)) / (\sum_i \mathit{length}(\mathit{visits\_d1}_i) * \sum_i \mathit{length}(\mathit{visits\_d2}_i))$ [‡], $(\sum_i \mathit{length}(\mathit{overlapping}_i)) / (\sum_i \mathit{length}(\mathit{visits\_d1}_i))$ [†]
Overlap and Location	$\{\mathit{present}(t) / \mathit{max\_present} \mid t \in \{\mathit{range}(\mathit{overlap\_start}, \mathit{overlap\_end}) \mid \mathit{visit} \in \mathit{overlapping}\}\}$ [*]
Individual and Location	$\{\mathit{present}(\mathit{visit\_start}) / \mathit{max\_present} \mid \mathit{visit} \in \mathit{visits\_d1}\}$ [*] [†], $\{\mathit{present}((\mathit{visit\_start} + \mathit{visit\_end})/2) / \mathit{max\_present} \mid \mathit{visit} \in \mathit{visits\_d1}\}$ [*] [†], $\{\mathit{present}(\mathit{visit\_end}) / \mathit{max\_present} \mid \mathit{visit} \in \mathit{visits\_d1}\}$ [*] [†]

4.2.3 Utilization Phase

After the regressor has been trained, it can be used to predict the social tie strengths between each pair of devices, and those tie strengths can be used to determine the aggregate social connectedness. At this point, the method takes only mobility features as input; the actual tie strengths are unnecessary.

As the first step in this phase, the method takes a time stamp as input, and uses it to determine which devices were present at the location at that moment in time. It again uses device time stamps to determine visit starts and ends. After doing so, the method determines the values of the mobility features that were given by the fea-

[*] This actually defines 7 statistical measures of a sequence of values, namely: their minimum, maximum, difference between minimum and maximum, mean, median, standard deviation and sum.

[†] While this feature is only shown for *d1*, it is generated for both devices, so it defines a pair of values. In order to make the order of those values independent from the order of the devices, the actual features are their maximum and minimum values.

[‡] This is the feature used in the baseline method.

ture selection algorithm for each pair of devices present. Then, these feature values are supplied to the regressor, which predicts the tie strength for each pair. Finally, the tie strengths are summed in order to obtain a measure of aggregate social connectedness. This value is then outputted.

Algorithm 3 shows the pseudo code for the utilization phase. Its inputs are the regressor generated during the initialization phase, the mobility trace, and a time stamp. Its output is the aggregate social tie strength for the given mobility trace and time stamp, as predicted by the regressor.

Algorithm 3: Utilization Phase

```
Data: <regressor r, mobility trace MT, timestamp t>
Result: aggregate social tie strength as
D = computePresentDevices(MT, t);
DP = computePairs(D);
PSC = [];
forall pair in DP do
    MF = computeMobilityFeatures(pair, MT);
    ps = predictTieStrength(regressor, MF);
    append(PSC,ps);
as = computeAggregateTieStrength(PSC);
return as;
```

5 Results

The methodology was evaluated based on a real-world dataset generated by Wi-Fi sensors that have been collecting probes from the University of Twente campus. Additionally, this data set was modified in various ways in order to analyze the methodology's sensitivity to different situations.

For each data set, we trained both the proposed and the baseline methodology, both generating their pair-wise social ties as output. This was repeated by using 10-folded cross-validation, and we compared their average coefficient of determination. For both the University of Twente campus data set and the modified data sets, we trained the regressors on multiple locations separately, in order to determine how well they perform under different environments.

The coefficient of determination (also known as R^2) is a measure of the proportion of the variance of the observed results that is explained by the predicted results. Its upper bound is 1, in which case all the observed results are predicted perfectly, and it does not have a lower bound. It is calculated as: $R^2 = 1 - \frac{\sum_i (f_i - y_i)^2}{\sum_i (y_i - \bar{y})^2}$, in which y_i is the ith predicted result and f_i is the ith observed result. The fraction's denominator is known as the total sum of squares, and is equivalent to the non-averaged variance of the observed data. The fraction's numerator is known as the residual sum of squares, and is equivalent to the non-averaged mean squared error (MSE). Since the value of R^2 is maximized by lowering the numerator, optimizing for R^2 will have the same result as minimizing the mean squared error.

The main reason that we chose to use the coefficient of determination as our error metric, is that it can be used regardless of the unit of the results. Many other metrics, such as the mean squared and absolute error, show the error in terms of commonly used units, such as meters when attempting to predict the lengths of objects. However, there are no commonly accepted units for social tie strength based on SSID list similarity, and showing the mean absolute error in terms of Psim-3 would be mostly meaningless for most people. For example, if we wanted to predict lengths, having a mean absolute error of 0.1 meters is meaningful for most people. In comparison, a mean absolute error of 0.01 Psim-3 likely has little meaning intuitively, and cannot be compared to other metrics of social tie strength. Therefore, we chose to sidestep this issue by using a metric that is dimensionless, and is therefore more generally applicable.

5.1 Dataset Descriptions

5.1.1 Real-world Dataset

The data set that will be used contains Wi-Fi data gathered by multiple sensors across the University Twente campus. These sensors have been gathering data since the start of 2016, as part of the Living Smart Campus project [47]. Each sensor receives Wi-Fi Request Probes from nearby devices, and stores related data in a central database.

For each request, the following items are stored:

- The timestamp of the request.
- The ID of the sensor that received the request.
- The anonymized MAC address of the user device. The address is anonymized by being salted and then hashed via the SHA-1 algorithm.
- The first 3 octets of the actual MAC address; these constitute the Organizationally Unique Identifier (OUI), indicating whether the address is possible randomized, or otherwise the device's manufacturer.
- The mean and maximum signal to noise ratio (SNR) during the request.
- The number of packets sent during the request.
- Each SSID that the device has broadcasted; these will later be anonymized in order to preserve user privacy.

Of these items, the timestamps, sensor IDs and MAC address are used to create the mobility trace, after the OUIs have been used to filter out possibly randomized MAC addresses. The SSIDs are solely used to infer the social tie strength between devices. As described in the background and methodology chapters, we use the Psim-3 metric for this purpose.

The dataset is gathered from 20 sensors for a period of 260 days. In this dataset, we found 2790703 distinct MAC addresses. However, a large proportion of these is likely randomized; after removing those addresses, only 281562 remain, approximately 10.09%. The sensors collected 130279931 probes in total, of which 126807946 were not from randomized sources, or 97.33%. Assuming that randomized and non-randomized sources produce the same number of probes on average, this means that 2.67% of the devices use randomized MAC addresses, and are responsible for 89.91% of addresses in the dataset. Figure 5.1 shows how the non-randomized MAC addresses are distributed among the top 25 SSIDs.

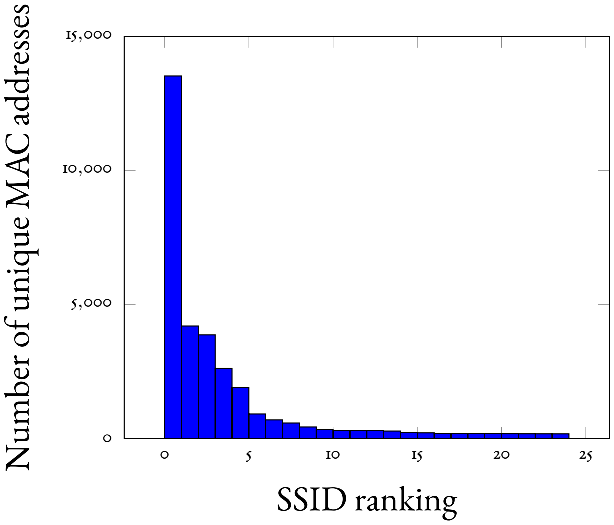

Figure 5.1: Total number of non-randomized MAC addresses per SSID

The regressors were trained separately on the data for three locations, which will be called location A through C. Location A is in the Vrijhof building, which main feature is the library, and often houses cultural performances. Location B is in the Sports Center, which contains sports facilities that are used by students and university employees. Location C is in the Spiegel, which is the administration center of the university. For each of these locations, the Wi-Fi sensors where placed in a coffee corner, where people can be expected to gather for social reasons. We chose these locations, because they represent very different types of location, and because we had the most data for these locations in the data set.

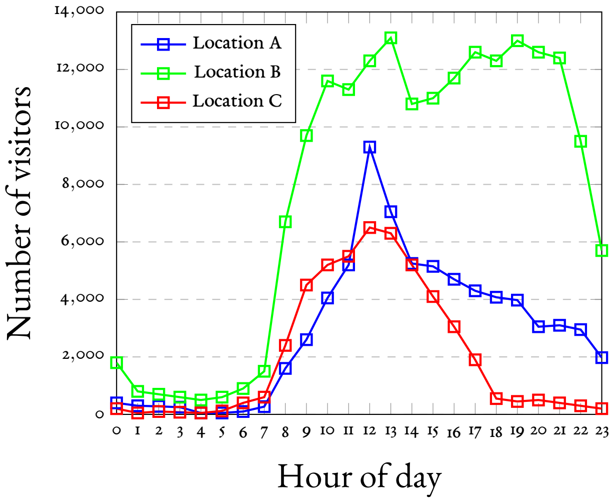

Figure 5.2: Total hourly number of visitors in period 02/2016 - 06/2016

Figure 5.2 shows the total number of visitors to each of these locations for different hourly timeslots. These numbers are lower than the actual number of visitors, because visitors without WiFi-enabled devices are not counted, and devices with randomized MAC addresses have been removed. It clearly shows that the Sports Center, location B, is more popular than the other two locations. Each of the locations has a peak around noon, while location B also has a peak around 7 PM, indicating that most people are at location A and C during the workday, while many people visit location B in the evening.

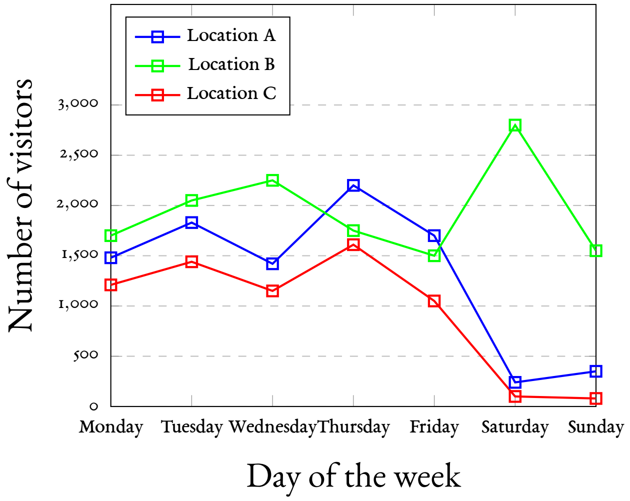

Figure 5.3: Total number of visitors at noon for different days of the week in period 02/2016 - 06/2016

Figure 5.3 shows the total number of visitors to each of the locations for different days of the week at noon. Again, location B seems more popular than the other two locations. Also, location A and C are busy from Monday to Friday, but not in the weekends, while location B is also busy in the weekend. This indicates that people visit A and C mostly during the work week, while they visit location B both during and outside the work week.

5.1.2 Modified Datasets

In addition to analyzing the method's performance on the real-world data, we do the same for several modifications of the original data set. The purpose of this is to determine how sensitive the method is to different circumstances, which gives an indication of how well it would perform in different environments. We considered generating synthetic data sets, which are based on simulated environments, as it would have allowed us to modify a larger variety of parameters. However, the main problem with doing so is that it would require a realistic simulation of social behavior, which to our knowledge has not yet been done with high accuracy.

Since we are trying to analyze the accuracy of a prediction of social ties, an inaccurate simulation of social relationships would interfere with that analysis. By modifying a data set based on the real world, we know for certain that the social ties are realistic, while still allowing us to analyze the effect of modifying a number of interesting parameters.

The modified data sets are generated by applying increasing degrees of the following adjustments:

- Adjustment 1: Remove probes from each device by decreasing frequency. For example, 50% of probes are removed by removing every second probe, and 75% is removed by only retaining the first, fifth, ninth, etcetera probe. This adjustment reflects an environment in which devices consistently broadcast fewer probes. This could be caused by different implementations of the 802.11 protocol or by using different protocols altogether.
- Adjustment 2: Remove samples before supplying them to the regressor. This change reflects a decrease of data available to the method, which could be caused by running the initialization phase for a lower amount of time, by placing the sensors in a location where few people gather, or by a larger number of people enabling MAC address randomization.
- Adjustment 3: Remove probes from each device probabilistically. This is similar to adjustment 1, but each probe is removed with a certain probability instead of by a fixed pattern. This adjustment simulates a situation in which fewer probes are received. This could be caused by a more noisy environment, by using sensors that are more susceptible to noise, or by making use of communication technology or protocols with less reliable transmission.

For each adjustment, a new set of samples was generated by applying the same process as that for the original data set, after which the same feature selection algorithm was used. Since adjustment 3 is randomized, performing the same adjustment may result in different levels of performance. In order to show a representative performance level, we performed the same adjustment 10 times and show the average coefficient of determination.

5.2 Analysis

In this section, we show and discuss the performance of the proposed and baseline methodologies on the data sets we have just discussed.

5.2.1 Real-world Dataset

Performance during feature selection

We start by examining the performance of both methodologies during the feature selection of the initialization phase. First, we look at the progression of performance over multiple rounds of the feature selection. We show this graph in order to demonstrate the difference in accuracy between the proposed and baseline method, ant to show how many features are used by the regressor. After this, we look at the impact of each feature class on the performance for the first two rounds of the feature selection. These graphs demonstrate the relative importance of different feature classes during different points.

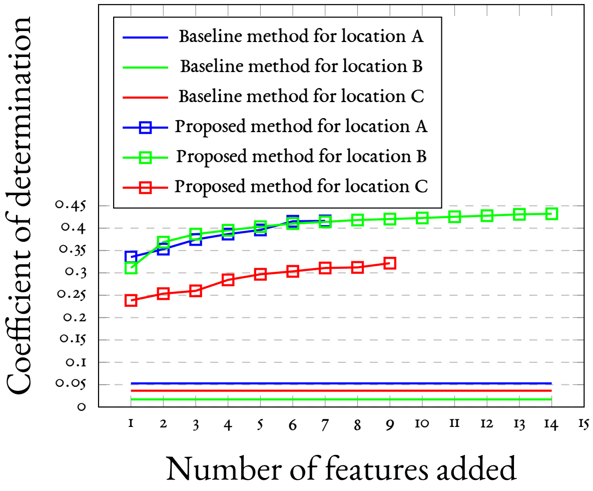

Figure 5.4: Progression of performance as features are added

Figure 5.4 shows the performance of the methodologies during the progression of the feature selection algorithm. As described, the algorithm keeps adding the feature with the highest performance (in terms of coefficient of determination) increase among all available unselected features, and stops when none of the features increases the performance. The graph shows that the proposed methodology reaches its optimal performance (for the given locations) by using 7 - 14 features, reaching a coefficient of determination between approximately 0.3 and 0.45. As expected, the baseline method (represented by a line as it uses a single feature) significantly underperforms, with a coefficient lower than 0.1. Also, the result for location A and B seems to be similar, while location B gets a lower score.

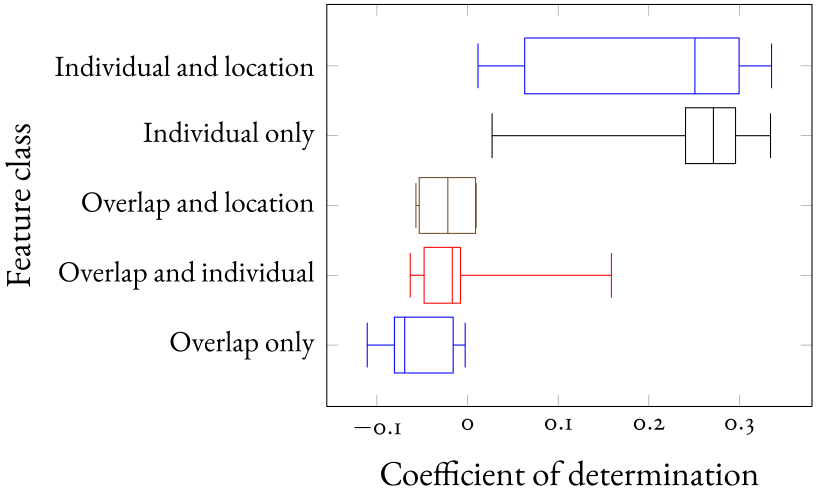

Figure 5.5: Performance after using a single feature in location A

Figure 5.5 shows the distribution of performances of the regressors generated during the first round of the feature selection, each of which only uses a single feature, for location A. It clearly shows that features based on individual mobility patterns outperform those based on overlapping visits. One possible reason for this is that ”Individual” features act as a filter for devices that are often present in the location without having strong social ties. Examples of these would be stationary devices such as vending machines, and devices owned by staff members.

The highest performing feature for this round was the higher sum of ratios between the number of devices present at the start of each device visit and the highest number of devices ever present in this location. This feature is one of the features defined by the first formula in the ”Individual and Location” row in table 4.1. If this feature has a high value, it would indicate that one of the devices is often present in the location when a lot of other devices are, which could mean that an overlapping visit with this device has little meaning. This would support the hypothesis that the feature acts as a filter.

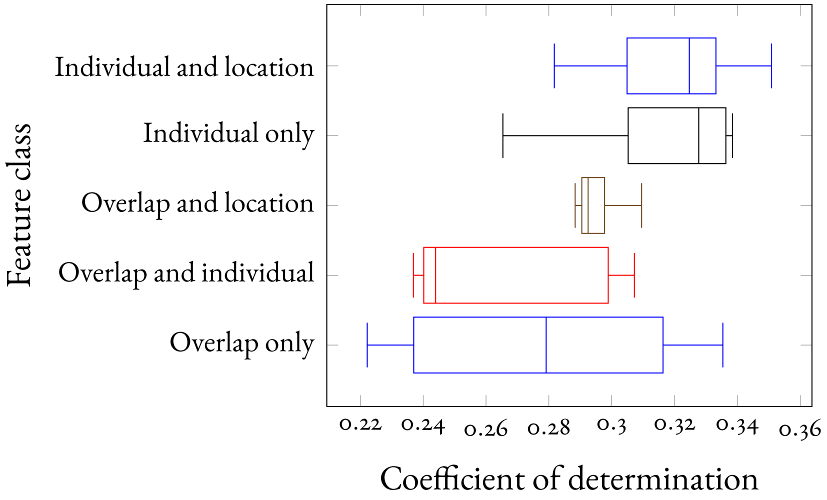

Figure 5.6: Performance after using two features in location A

Figure 5.6 shows the same distributions, but for the second feature selection round, in which each regressor uses the feature from the previous round and a newly selected feature. The main difference with the previous graph is that the feature classes related to overlap seem to improve their performance. For each of the locations, we found that features from the overlap classes are selected in later rounds. This is consistent with the idea that individual-related features could act as filters: once devices without social ties are filtered out, overlap features could indicate which social ties are stronger.

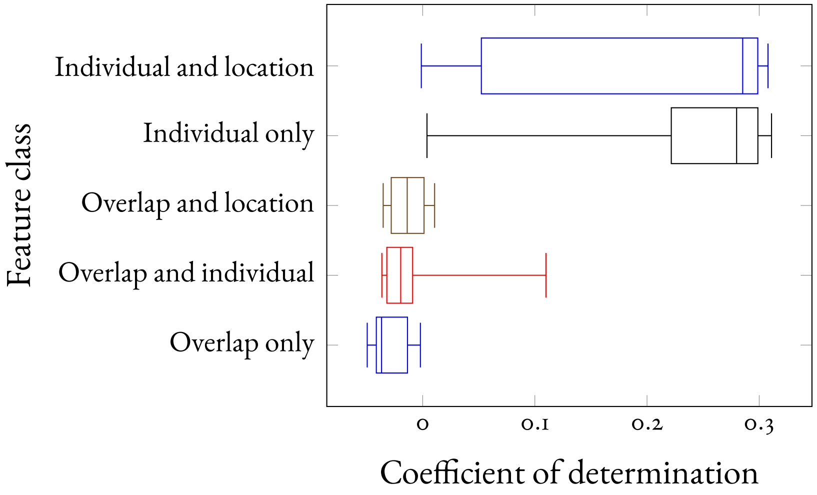

Figure 5.7: Performance after using a single feature in location B

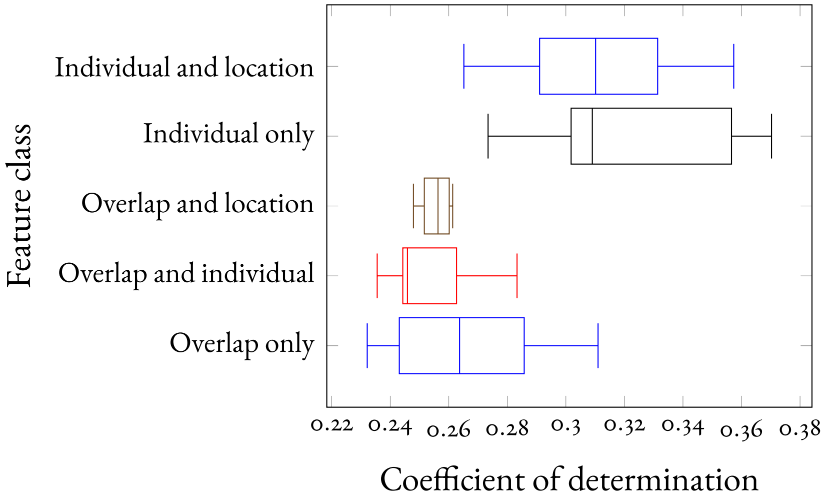

Figure 5.8: Performance after using two features in location B

Figure 5.7 and 5.8 show a similar pattern to the two preceding graphs. The main difference is that individual-related features seem to more strongly outperform overlap features, even during the second round. This could be caused by a higher degree of non-social devices in this location.

The highest performing feature for this round was the maximum visit length to this location of either device. If this feature has a high value, it would indicate that one of the devices is present in this location for long periods of time. As with the best-performing feature for location A, this could mean that an overlapping visit with this device has little meaning.

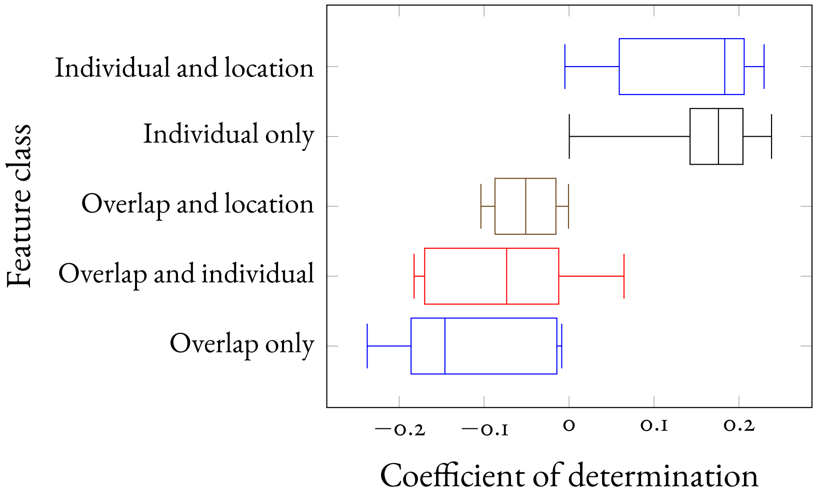

Figure 5.9: Performance after using a single feature in location C

Again, the same pattern holds for the distributions in graph 5.9 and 5.10. The main difference with the previous graphs is that some of the overlap features already outperform individual-related features in the second round.

The highest performing feature for this round was the lower sum of visit lengths to this location between both devices. Like the previous two best-performing features, a high value for this feature indicates that at least one of the devices is present in this location for long periods of time and that an overlapping visit between the pair of devices has little meaning socially.

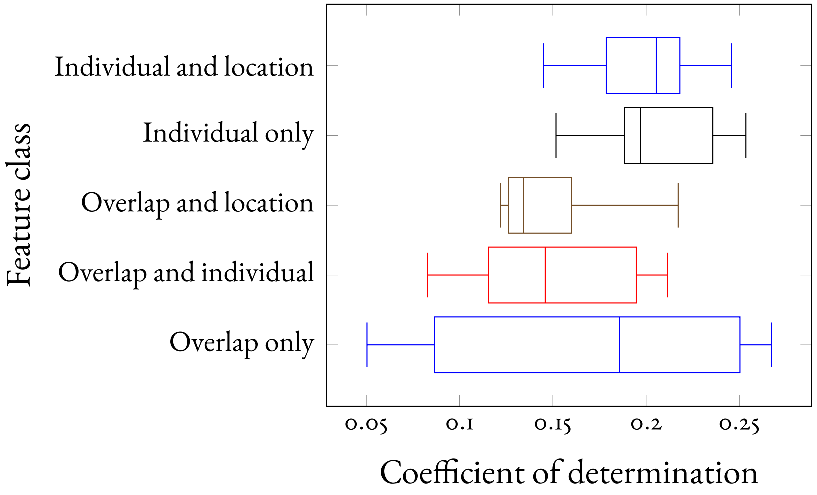

Figure 5.10: Performance after using two features in location C

From the shown graphs, we can conclude that features related to the devices themselves have a greater impact on predicting social tie strengths than those related to overlap. Because of this, the accuracy of the proposed method is already higher than that of the baseline method in its first feature selection round. We can also conclude that once a feature related to devices has been selected, the performance of overlap-related features increases significantly. One possible explanation for this is that device-related features act as 'filters', and are used to determine static devices that are often present in the location, but do not have strong social ties with many visiting devices.

In order to compare the algorithm's performance for multiple locations, we trained regressors for all 20 locations, while using the features from our best performing location, location B. Figure 5.11 shows the resulting scores, set out against the number of unique MAC addresses that visited the location. The figure makes clear that there is a strong correlation between location popularity and the accuracy of the proposed method, but not with the accuracy of the baseline method. It also shows that the regressors performs well when a static feature set is used, instead of using the feature selection algorithm.

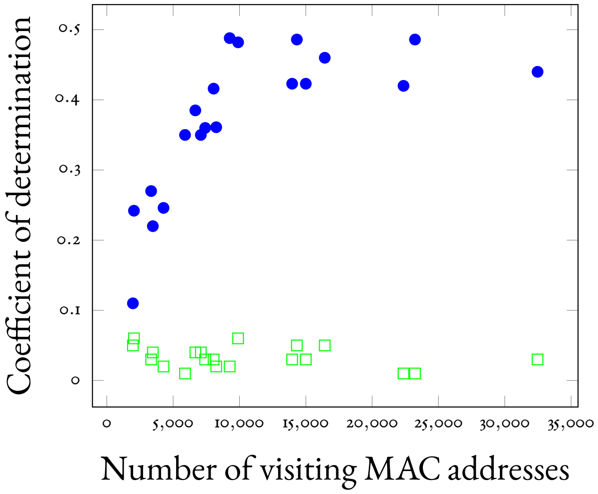

Figure 5.11: Performance for multiple locations set out against location popularity

Aggregate social tie strengths

In this subsection, we look at the aggregate social tie strengths over time for different locations. We compare the actual tie strengths (obtained from the data set) to the strengths predicted by the proposed and baseline methods. These graphs show both the progression of social connectedness over the day and week, and give an idea of the difference in accuracy between the baseline and proposed methods.

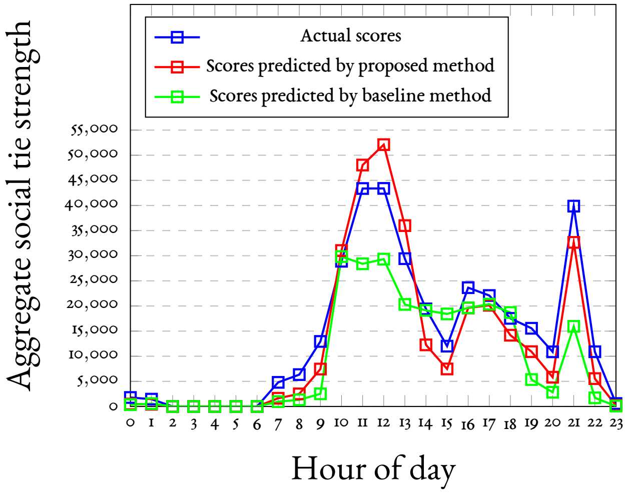

Figure 5.12: Aggregate social tie strength by hour of day for location A

Figures 5.12 and 5.13 show the aggregated social tie strengths for different hours of the day for two locations, as computed in the utilization phase. It shows the scores computed from the actual social ties, and those computed from the scores predicted by the proposed method and the baseline method. For both locations, the scores from the predicted method are generally closer to the actual scores than those of the baseline, and shape of the proposed method graph also resembles that of the actual graph more closely. This result is expected, because the pair-wise tie strengths were also better predicted by the proposed method. One interesting fact is that the graphs for location A have a peak around 9PM, while there is no such peak in the number of visitors at the same time. This may be explained by the fact that there are sometimes cultural events at location A in the evening; it may be possible that visitors to these events are more strongly connected socially than regular visitors.

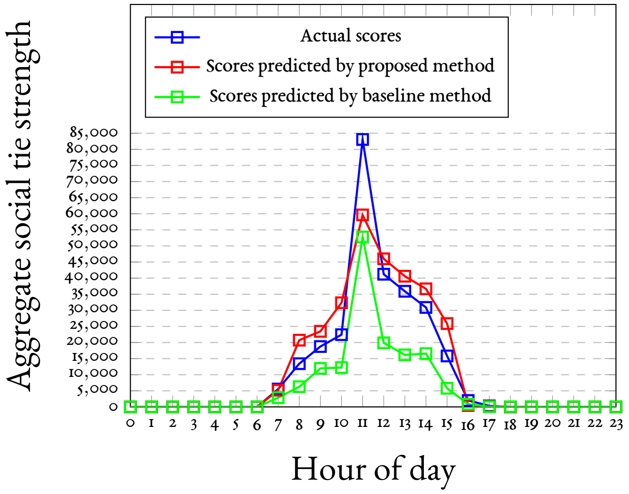

Figure 5.13: Aggregate social tie strength by hour of day for location C

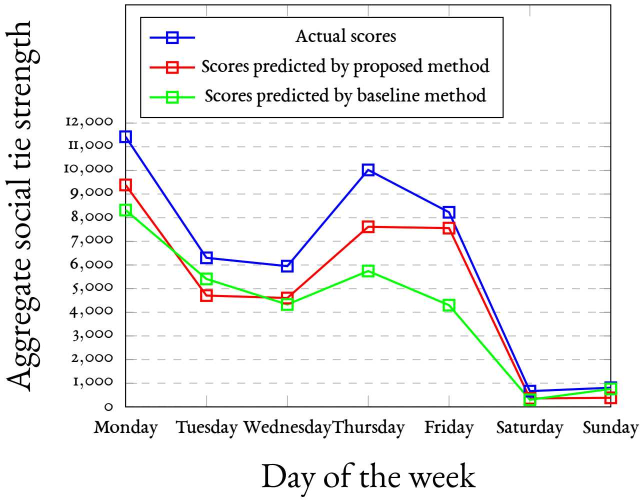

Figure 5.14: Aggregate social tie strength by day of the week for location A

Figures 5.14, 5.15 and 5.16 show the same aggregated scores as the previous two figures, but for different days of the week. Again, the scores for the proposed methodology follow the real scores more closely than those for the baseline method, as expected. Like the graphs on visitor frequency, location A and C are more busy on weekdays than on the weekend, while location B is also busy in the weekend. However, the pattern in aggregate social tie strength does not completely match that of the number of visitors. For example, the tie strength on Monday for location A and B is higher than that on Tuesday, while there are more visitors. The reason for this is yet unclear.

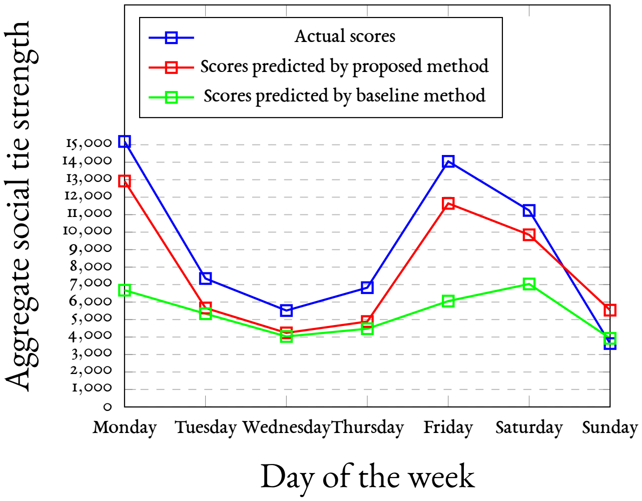

Figure 5.15: Aggregate social tie strength by day of the week for location B

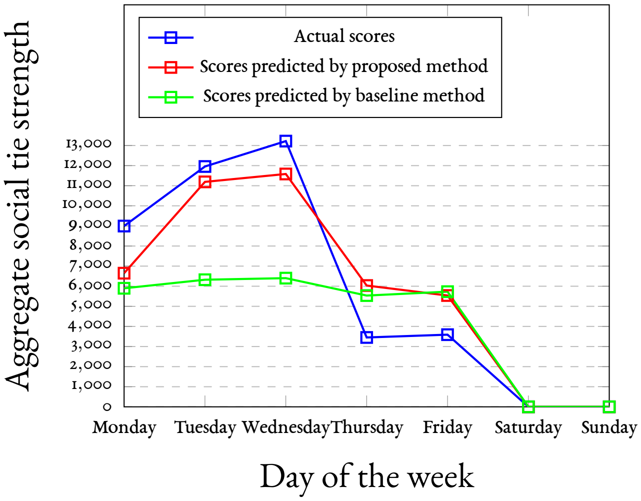

Figure 5.16: Aggregate social tie strength by day of the week for location C

5.2.2 Modified Datasets

The following figures show the performance of the proposed and baseline methods on the modified data sets we described earlier. As before, we first show the performance during the feature selection. Then, in the subsequent graphs, we increase the degree of each adjustment and compare the method's resulting accuracy. The main purpose of these graphs is to demonstrate the robustness of the proposed method under unfavorable environmental circumstances.

Performance during feature selection

Figure 5.17 shows the progression of performance for location A while using the feature selection algorithm for both the unadjusted data set and the data sets after adjustment 1 (50% of probes removed by decreasing frequency), adjustment 2 (50% of samples removed) and adjustment 3 (50% of probes removed probabilistically). The figure shows that the proposed method performs better than the baseline method on each of the data sets, and that each of the adjustments decreases the performance slightly for both methods.

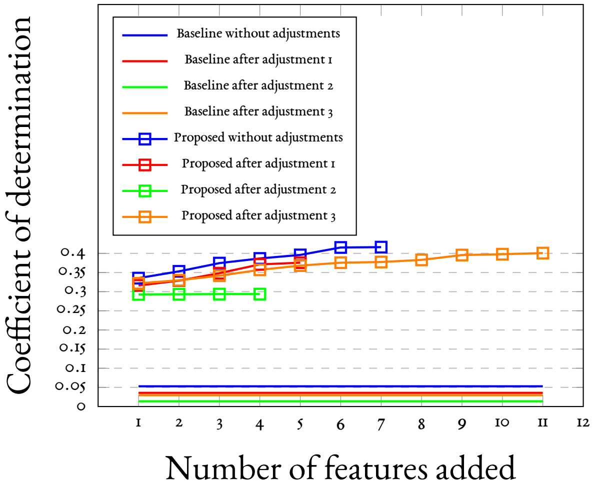

Figure 5.17: Progression of performance for location A after data set adjustments

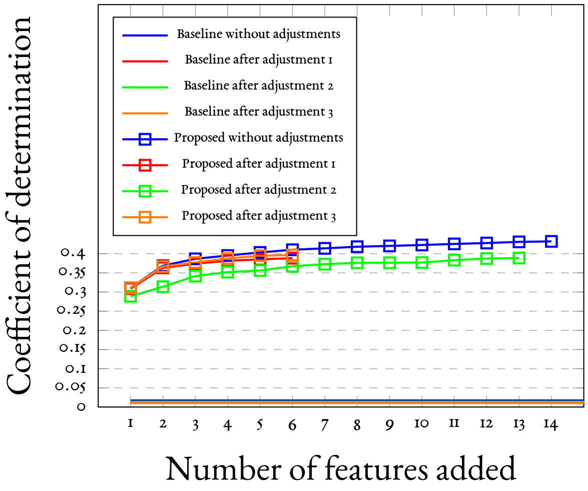

Figure 5.18: Progression of performance for location B after data set adjustments

Figures 5.18 and 5.19 show the same data, but for location B and C. For these locations also, the proposed method performs better than the baseline method, and both perform worse after the data set adjustments. It is also clear that adjustment 2 has the largest negative impact of the three adjustments for the proposed method, but not in each location for the baseline method.

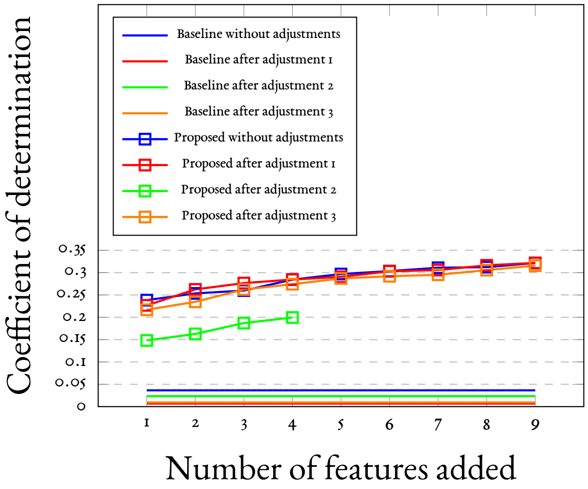

Figure 5.19: Progression of performance for location C after data set adjustments

Final performance for different degrees of adjustment

The next three figures show the final performance of the proposed method, which is the accuracy of the regressor after the feature selection algorithm has completed. Each figure shows this performance for different degrees of one of the three adjustment types, for each of the locations. For example, figure 5.20 shows the proposed and baseline performances when 0% (unadjusted), 50%, 75%, etc. of the probes have been removed by decreasing frequency. As expected, the proposed method performs better than the baseline in every case. Also, higher degrees of the adjustment decrease performance of both methods in nearly every case. Secondly, the performance decreases more quickly at the last degrees of adjustment for location A and C than for location B. This may be caused by the fact that these locations have fewer visitors than location B, so removing over 95% of the probes may make it impossible to find any meaningful overlapping visits.

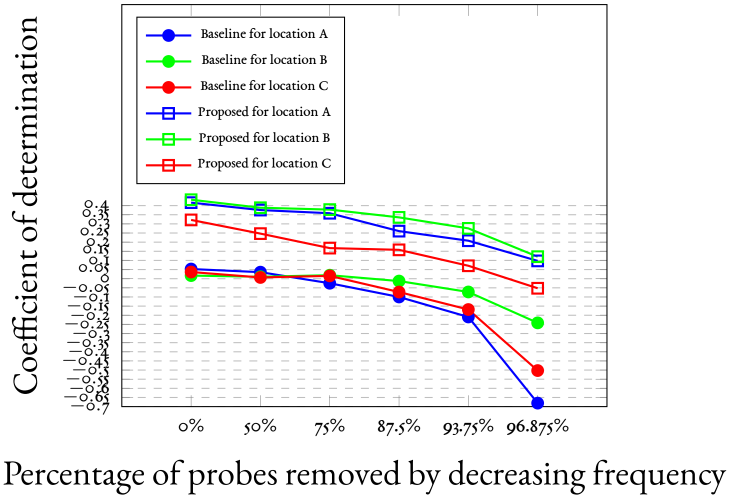

Figure 5.20: Final performance for different degrees of adjustment 1

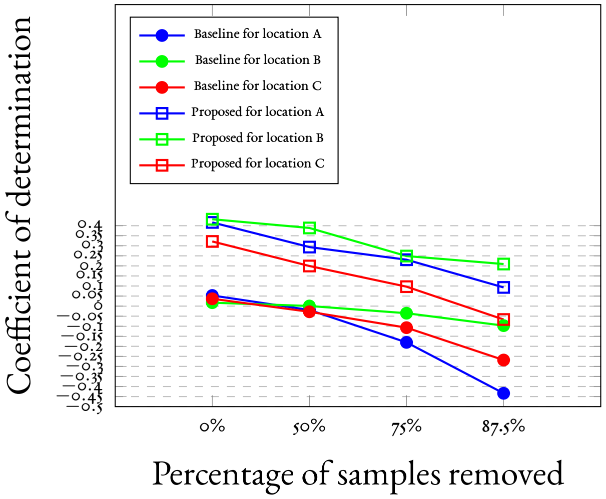

Figure 5.21: Final performance for different degrees of adjustment 2

The performances for different degrees of adjustment 2 and 3 are shown in figures 5.21 and 5.22. The general observations as those for figure 5.20 hold. Each of the figures shows that a large amount of input data can be removed before the proposed method performs worse than the baseline method on the unadjusted data. This suggests that the proposed method would also work well in locations with negative external circumstances such as noise.

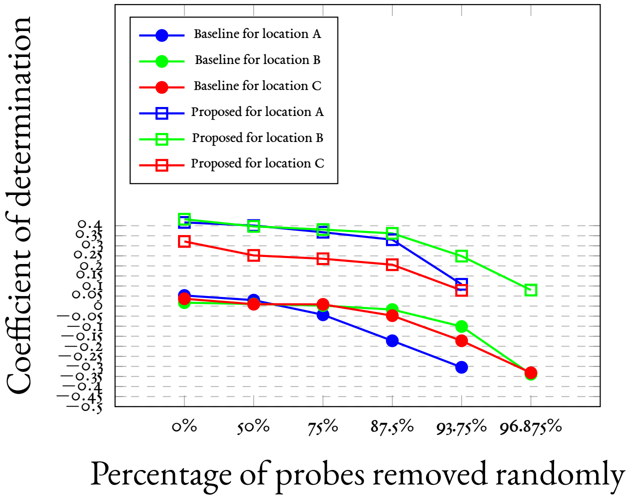

Figure 5.22: Final performance for different degrees of adjustment 3

From these graphs, we can conclude that the proposed method continues to do well even in very unfavorable circumstances, as simulated by the data set adjustments. In each examined case, the proposed method continues to outperform the baseline method. Specifically, the proposed method has an accuracy similar to the unadjusted data set when 75% of probes are removed (either by decreasing frequency or probabilistically), and its performance decreases to a slower degree than that of the baseline method. From this, we conclude that the proposed method is more robust than the baseline.

6
Conclusion

In this chapter, we answer the research questions we posed at the start of the thesis, and propose avenues for future research.

6.1 Research Questions

In this study, we aimed to answer the following research question: *Is it possible to use patterns in device mobility data to infer the social connectedness for a given location and moment in time?*. In order to do so, we posed two sub research questions, which we can now answer.

The first sub question was: *Which metrics that are correlated with social tie strength* *are observable from Wi-Fi data?* This question was answered by a literature study on the Wi-Fi protocol and similarity metrics. We found that multiple previous studies had used the set of SSIDs broadcasted by each device as a basis to compute a similarity score between each pair of devices, which is correlated with social tie strength. There is, however, a large variety in the similarity metrics used by these studies, including the Jaccard index, TF-IDF, Adamic-Adar, and a number of modifications of these more general metrics. We selected the metric most appropriate to our application based on an evaluation of their correlation with social tie strength, while taking into account the properties that were required for the application. In conclusion, we selected the Psim-3 metric (a modification of Adamic-Adar), as it has one of the highest correlations with social linkage, and is sensitive to increased identifier rarity and count.

The second sub question was: *Is it possible to accurately predict these social tie strength metrics from general (not Wi-Fi specific) device mobility data?* This question was answered by implementing a new methodology and evaluation its performance on a number of data sets. This method works by selecting a number of mobility features from a large set of possible features, and by learning the relationship between these features and social tie strengths by using a regressor. Once this regressor has been trained, it can be used to predict social tie strengths by supplying only a mobility trace. The method was evaluated based on a number of real-world data sets and a number of modified versions of those data sets, in order to determine the method's sensitivity to various parameters. Furthermore, we compared this performance to that of a state of the art baseline method. We found that the proposed method can explain about 30% to 45% of the variance in social tie strength in the unmodified data sets. For each of the modified data sets, we found that the proposed method's performance degraded slowly, signifying a robustness to complicating external circumstances. In each case, the proposed methodology performed significantly better that the baseline methodology.

In conclusion, it is indeed possible to infer social connectedness from only device mobility data, to a large degree. This is consistent with our hypothesis, which we based on the well-supported sociological theory of homophily.

6.2 Future Work

We now look at a number of possible future directions for research, given these conclusions.

The proposed method works by selecting a subset of features by computing a larger set and using a greedy feature selection algorithm. The performance of the method can therefore be expected to be highly dependent on both the computed features and the selection algorithm. Future studies could investigate other mobility features that may be correlated with social tie strength, like combinations of features previously discussed, or features that have not yet been described in current research. They could also use different regressors, or use feature selection algorithms in order to determine better combinations of features.

Additionally, the method was evaluated by using a metric of SSID list similarity. While existing research supports a degree of correlation between this similarity and social linkage, the correlation is not 100%. Also, this method of evaluation is not possible for many other sources of mobility data. Therefore, a future study could attempt to create a method similar to the proposed one that uses a different measure that may be more generally available or has a stronger correlation with actual social tie strength.

Thirdly, the evaluation showed that the proposed method's performance varies when applied in different locations. This would suggest that for some locations, there is a stronger correlation between mobility and social relationships than others. Future research could try to detect this correlation strength more systematically, or create a method that is more robust to this variation.

Finally, we suggested a number of possible applications at the start of the thesis, such as determining a statistical measure of social connection in care facilities, improving MANET routing algorithms and enabling social user applications. Other studies could incorporate the proposed methodology and evaluate its usefulness and performance for these and other applications.

References

[1] Adamic, L. A. & Adar, E. (2003). Friends and neighbors on the web. *Social networks*, 25(3), 211–230.

[2] Ananthanarayanan, G. & Stoica, I. (2009). Blue-fi: enhancing wi-fi performance using bluetooth signals. In *Proceedings of the 7th international conference on Mobile systems, applications, and services* (pp. 249–262).: ACM.

[3] Baratchi, M., Heijenk, G., & Van Steen, M. (2016). Spaceprint: a mobility-based fingerprinting scheme for public spaces.

[4] Barbera, M. V., Epasto, A., Mei, A., Perta, V. C., & Stefa, J. (2013). Signals from the crowd: uncovering social relationships through smartphone probes. In *Proceedings of the 2013 conference on Internet measurement conference* (pp. 265–276).: ACM.

[5] Berkman, L. F., Kawachi, I., & Glymour, M. M. (2014). *Social epidemiology*. Oxford University Press.

[6] Bonné, B., Barzan, A., Quax, P., & Lamotte, W. (2013). Wifipi: Involuntary tracking of visitors at mass events. In *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2013 IEEE 14th International Symposium and Workshops on a* (pp. 1–6).: IEEE.

[7] Bonné, B., Quax, P., & Lamotte, W. (2014). Your mobile phone is a traitor!–raising awareness on ubiquitous privacy issues with sasquatch.

[8] Cacioppo, J. T. & Cacioppo, S. (2014). Social relationships and health: The toxic effects of perceived social isolation. *Social and personality psychology compass*, 8(2), 58–72.

[9] Calhoun, P., Cisco Systems, Montemurro, M., Research in Motion, Stanley, D., & Aruba Networks (2009). *Control and Provisioning of Wireless Ac-*

cess Points (CAPWAP) Protocol Binding for IEEE 802.11. RFC 5416, RFC Editor.

[10] Cheng, N., Mohapatra, P., Cunche, M., Kaafar, M. A., Boreli, R., & Krishnamurthy, S. (2012). Inferring user relationship from hidden information in wlans. In *MILCOM 2012-2012 IEEE Military Communications Conference* (pp. 1–6).: IEEE.

[11] Chernyshev, M., Valli, C., & Hannay, P. (2015). 802.11 tracking and surveillance-a forensic perspective. In *Proceedings of the International Conference on Security and Management (SAM)* (pp. 349).: The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

[12] Christakis, N. A. & Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *New England journal of medicine*, 357(4), 370–379.

[13] Cisco Systems (2016). 10th annual cisco visual networking index (vni) mobile forecast projects 70 percent of global population will be mobile users. https://newsroom.cisco.com/press-release-content?type=webcontent&articleId=1741352. [Online; accessed 21-November-2016].

[14] Cunche, M. (2014). I know your mac address: Targeted tracking of individual using wi-fi. *Journal of Computer Virology and Hacking Techniques*, 10(4), 219–227.

[15] Cunche, M., Kaafar, M.-A., & Boreli, R. (2014). Linking wireless devices using information contained in wi-fi probe requests. *Pervasive and Mobile Computing*, 11, 56–69.

[16] Di Luzio, A., Mei, A., & Stefa, J. (2016). Mind your probes: De-anonymization of large crowds through smartphone wifi probe requests. In *INFOCOM*: IEEE.

[17] Do, T. M. T. & Gatica-Perez, D. (2012). Contextual conditional models for smartphone-based human mobility prediction. In *Proceedings of the 2012 ACM conference on ubiquitous computing* (pp. 163–172).: ACM.

[18] Duong, T. V. T. & Tran, D. Q. (2012). An effective approach for mobility prediction in wireless network based on temporal weighted mobility rule. *International Journal of Computer Science and Telecommunications*, 3(2), 29–36.

[19] Eisenberger, N. I. & Cole, S. W. (2012). Social neuroscience and health: neurophysiological mechanisms linking social ties with physical health. *Nature neuroscience*, 15(5), 669–674.

[20] Holt-Lunstad, J., Smith, T. B., Baker, M., Harris, T., & Stephenson, D. (2015). Loneliness and social isolation as risk factors for mortality a meta-analytic review. *Perspectives on Psychological Science*, 10(2), 227–237.

[21] Hui, P., Crowcroft, J., & Yoneki, E. (2011). Bubble rap: Social-based forwarding in delay-tolerant networks. *IEEE Transactions on Mobile Computing*, 10(11), 1576–1589.

[22] IEEE (2012). Ieee standard for information technology–telecommunications and information exchange between systems local and metropolitan area networks–specific requirements part 11: Wireless lan medium access control (mac) and physical layer (phy) specifications.

[23] IEEE-SA (2011). Standard group mac addresses: A tutorial guide.

[24] Jylhä, M. & Aro, S. (1989). Social ties and survival among the elderly in tampere, finland. *International Journal of Epidemiology*, 18(1), 158–164.

[25] Kawachi, I. & Berkman, L. F. (2001). Social ties and mental health. *Journal of Urban health*, 78(3), 458–467.

[26] Kim, M. & Kotz, D. (2005). Modeling users' mobility among wifi access points. In *Papers presented at the 2005 workshop on Wireless traffic measurements and modeling* (pp. 19–24).: USENIX Association.

[27] Langley, P. et al. (1994). Selection of relevant features in machine learning. In *Proceedings of the AAAI Fall symposium on relevance*, volume 184 (pp. 245–271).

[28] Liben-Nowell, D. & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7), 1019–1031.

[29] Lindqvist, J., Aura, T., Danezis, G., Koponen, T., Myllyniemi, A., Mäki, J., & Roe, M. (2009). Privacy-preserving 802.11 access-point discovery. In *Proceedings of the second ACM conference on Wireless network security* (pp. 123–130).: ACM.

[30] Mashhadi, A., Vanderhulst, G., Acer, U. G., & Kawsar, F. (2015). An autonomous reputation framework for physical locations based on wifi sig-

nals. In *Proceedings of the 2nd workshop on Workshop on Physical Analytics* (pp. 43–46).: ACM.

[31] Mashhadi, A. J., Mokhtar, S. B., & Capra, L. (2009). Habit: Leveraging human mobility and social network for efficient content dissemination in delay tolerant networks. In *World of Wireless, Mobile and Multimedia Networks & Workshops, 2009. WoWMoM 2009. IEEE International Symposium on a* (pp. 1–6).: IEEE.

[32] McGee, J., Caverlee, J., & Cheng, Z. (2013). Location prediction in social media based on tie strength. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (pp. 459–468).: ACM.

[33] McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, (pp. 415–444).

[34] Metzler, D., Bernstein, Y., Croft, W. B., Moffat, A., & Zobel, J. (2005). Similarity measures for tracking information flow. In *Proceedings of the 14th ACM international conference on Information and knowledge management* (pp. 517–524).: ACM.

[35] Mishra, A., Shin, M., & Arbaugh, W. (2003). An empirical analysis of the ieee 802.11 mac layer handoff process. *ACM SIGCOMM Computer Communication Review*, 33(2), 93–102.

[36] Misra, B. (2014). ios8 mac randomization – analyzed! http://blog.mojonetworks.com/ios8-mac-randomization-analyzed/. [Online; accessed 21-November-2016].

[37] Musa, A. & Eriksson, J. (2012). Tracking unmodified smartphones using wi-fi monitors. In *Proceedings of the 10th ACM conference on embedded network sensor systems* (pp. 281–294).: ACM.

[38] National Institute on Aging (2015). Humanity's aging. https://www.nia.nih.gov/research/publication/global-health-and-aging/humanitys-aging. [Online; accessed 11-November-2016].

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

[40] Perissinotto, C. M., Cenzer, I. S., & Covinsky, K. E. (2012). Loneliness in older persons: a predictor of functional decline and death. *Archives of internal medicine*, 172(14), 1078–1084.

[41] Scellato, S., Noulas, A., & Mascolo, C. (2011). Exploiting place features in link prediction on location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1046–1054).: ACM.

[42] Schauer, L., Werner, M., & Marcus, P. (2014). Estimating crowd densities and pedestrian flows using wi-fi and bluetooth. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services* (pp. 171–177).: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[43] Seeman, T. E. (1996). Social ties and health: The benefits of social integration. *Annals of epidemiology*, 6(5), 442–451.

[44] Song, L., Kotz, D., Jain, R., & He, X. (2004). Evaluating location predictors with extensive wi-fi mobility data. In *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies*, volume 2 (pp. 1414–1424).: IEEE.

[45] Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1), 11–21.

[46] Storper, M. & Venables, A. J. (2004). Buzz: face-to-face contact and the urban economy. *Journal of economic geography*, 4(4), 351–370.

[47] University of Twente (2016). Privacy-preserved crowd monitoring. https://www.utwente.nl/en/organization/news-agenda/special/2016/living-smart-campus/projects/privacy-preserved-crowd-monitoring/. [Online; accessed 21-November-2016].

[48] Vanhoef, M., Matte, C., Cunche, M., Cardoso, L. S., & Piessens, F. (2016). Why mac address randomization is not enough: An analysis of wi-fi network discovery mechanisms. In *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security* (pp. 413–424).: ACM.

[49] Vu, L., Do, Q., & Nahrstedt, K. (2011). Jyotish: A novel framework for constructing predictive model of people movement from joint wifi/bluetooth

trace. In *Pervasive Computing and Communications (PerCom), 2011 IEEE International Conference on* (pp. 54–62).: IEEE.

[50] Yavaş, G., Katsaros, D., Ulusoy, Ö., & Manolopoulos, Y. (2005). A data mining approach for location prediction in mobile environments. *Data & Knowledge Engineering*, 54(2), 121–146.

[51] Zemel, R. S. & Pitassi, T. (2001). A gradient-based boosting algorithm for regression problems. *Advances in neural information processing systems*, (pp. 696–702).