Analytics of Human Presence and Movement Behaviour Within Specific Environments

Muawya Habib Sarnoub Eldaw

December 2019

A Thesis Submitted to
Birkbeck, University of London
in Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

Department of Computer Science & Information Systems
Birkbeck
University of London

Declaration

This thesis is the result of my own work, except where explicitly acknowledged in the text.

Muawya Habib Sarnoub Eldaw _____

December 18, 2019

Abstract

The vast amounts of detailed information, generated by Wi-Fi and other mobile communication technologies, provide an invaluable opportunity to study different aspects of presence and movement behaviours of people within a given environment; for example, a university campus, an organisation office complex, or a city centre. Utilising such data, this thesis studies three main aspects of the human presence and movement behaviours: spatio-temporal movement (where and when do people move), user identification (how to uniquely identify people from their presence and movement historical records), and social grouping (how do people interact). Previous research works have predominantly studied two out of these three aspects, at most. Conversely, we investigate all three aspects in order to develop a coherent view of the human presence and movement behaviour within selected environments. More specifically, we create stochastic models for movement prediction and user identification. We also devise a set of clustering models for the detection of the social groups within a given environment.

The thesis makes the following contributions:

1. Proposes a family of predictive models that allows for inference of locations though a collaborative mechanism which does not require the profiling of individual users. These prediction models utilise suffix trees as their core underlying data structure, where predictions about a specific individual are computed over an aggregate model incorporating the collective record of observed behaviours of multiple users.

2. Defines a mobility fingerprint as a profile constructed from the users historical mobility traces. The proposed method for constructing such a profile is a principled and scalable implementation of a variable length Markov model based on n-grams.

3. Proposes density-based clustering methods that discover social groups by analysing activity traces of mobile users as they move around, from one location to another, within an observed environment.

We utilise two large collections of mobility traces: a GPS data set from Nokia and an Eduroam network log from Birkbeck, University of London, for the evaluation of the proposed models reported herein.

Publications

The following publications by the author are related to this thesis:

⟨+⟩ Eldaw, Muawya Habib Sarnoub, Mark Levene, and George Roussos. "Collective suffix tree-based models for location prediction." In Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication, pp. 441-450. ACM, Zurich, 2013, Zurich, Switzerland.

⟨+⟩ Eldaw, Muawya Habib Sarnoub, Mark Levene, and George Roussos. "Poster: Constructing a Unique Profile for Mobile User Identification in Location Recommendation Systems." In Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services, pp. 479-479. ACM, 2015, Florence, Italy.

⟨+⟩ Eldaw, Muawya H. Sarnoub, Mark Levene, and George Roussos. "Presence analytics: Discovering meaningful patterns about human presence using wlan digital imprints." In Proceedings of the International Conference on Internet of things and Cloud Computing, p. 53. ACM, 2016, Cambridge, UK.

⟨+⟩ Eldaw, Habib Sarnoub, Mark Levene, and George Roussos. "Presence analytics: density-based social clustering for mobile users." In ICETE 2016: Proceedings of the 13th International Joint Conference on e-Business and Telecommunications table of contents, pp. 52-62. SCITEPRESS-Science and Technology Publications, Lda, 2016, Lisbon, Portugal.

⟨+⟩ Eldaw, Muawya Habib Sarnoub, Mark Levene, and George Roussos. "Social-DBSCAN: A Presence Analytics Approach for Mobile Users Social Clustering." In International Conference on E-Business and Telecommunications, pp. 381-400. Springer, Cham, 2016.

⟨+⟩ Eldaw, Muawya H. Sarnoub, Mark Levene, and George Roussos. "Presence analytics: Detecting classroom-based social patterns using WLAN traces." In 2017 Intelligent Systems Conference (IntelliSys), pp. 346-353. IEEE, 2017, London, UK.

⟨+⟩ Eldaw, Muawya Habib Sarnoub, Mark Levene, and George Roussos. "Presence analytics: making sense of human social presence within a learning environment." In 2018 IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT), pp. 174-183. IEEE, 2018, Zurich, Switzerland.

Acknowledgements

My thanks to ...

Family

My parents, Habib Sarnoub and Habiba Saad, your strength has always inspired me and I am eternally indebted to you for everything you did for me.

My wife Nada Rahhal, my son Alhassan, my daughters Rayyan, Maryam and Marya, thank you for being there whenever I needed your love and support.

My brothers and sister, Hatim, Mohamed and Sara for your unwavering support and encouragement.

Colleagues and Friends

My colleagues in the research lab at the Department of Computer Science and Information Systems at Birkbeck, and my friends elsewhere for their kind support during a long but enjoyable process.

Supervisors

Professor George Roussos, for your support and valuable feedback. My thanks are also extended to Professor George Loizou for the invaluable guidance concerning the content and proof reading of this thesis.

Finally, my academic supervisor during MSc and PhD, Professor Mark Levene. Thank you for your support and guidance over the past years. I learned a lot from you and I will always be grateful to you.

Contents

List of Abbreviations

AP Access Point.

CCDF Complementary Cumulative Distribution Function.
CMPL Collective Most Popular Location.
CohCoff Coherence Coefficient.
CST Collective Suffix Tree.
CTSP Collective Time-spent Predictor.

DBMS Database Management System.
DBSCAN Density-based Spatial Clustering of Applications with Noise.
DNA Deoxyribonucleic Acid.

EDUM EDUca-tion Measurement.
Eduroam Education Roaming.
EPFL École Polytechnique Fédérale de Lausanne.

GDPR General Data Protection Regulation.
GPS Global Positioning System.

HM Hit and Miss Score.

IC Identification Coefficient.
Idiap Idiap Research Institute.
ITS Birkbeck's IT Services.

JSD Jensen-Shannon Divergence.

KLD Kullback-Leibler Divergence.

LDCC Lausanne Data Collection Campaign.
LHC Large Hadron Collider.

MAC Media Access Control Address.
MAE Mean Absolute Error.
MDC Mobile Data Challenge.
MinSGroupSize Minimum Size of a Social Group.
MLE Maximum Likelihood Estimation.
MPL Most Popular Location.

QDA Quadratic Discriminant Analysis.

RMSE Root Mean Square Error.
RSSI Receive Signal Strength Indicators.

SOAS School of Oriental and African Studies.
SPM The Social Presence Model.
ST Suffix Tree.
SVM Support Vector Machine.

TSP Time-spent Predictor.

UCL University College London.

WJaccard Weighted Jaccard.
WLAN Wireless Area Network.

XML eXtensible Markup Language.

List of Symbols

Symbol	Meaning
A, B	Denote the training data subsets.
α	Denotes the percentage of the data used for training, and thus, 100-α denotes the percentage used for testing.
T, τ	A mobility trail made of a ordered sequence of locations.
T_{l}	denotes the set of all trails of the observed landmark l.
N	A finite set of movement trails.
l	A single location or landmark.
S	A suffix tree.
u	A single user.
U	Database of users.
L	The set of locations.
$\vartheta_{T,n}$	A set comprises all n-pairs generated by using the elements of T.
$\beta_{T,n}$	A set that contains all the possible *n-grams* of a mobility trail T.
$B_{u,n}$	A set of all the *n-grams* from all the mobility trails a user u.
F, M	A set of fingerprints
f_u	A fingerprint of an observed user u.
$\gamma(i)$	A function that returns the number of occurrences of i.
θ	A ranking threshold value that maximises, pointwise, the *area under the precision and recall curve.*
λ	A similarity threshold value.
p, q, r	An m-dimensional point representing a user's visits to locations in L.
W	The combined records of visits of a given day over the 11 weeks academic term
σ	A user-defined threshold value.
n	The number of items in a *set of items*, and thus it means the size of such a set.
v	A user's visit, to a given location, within a time interval t.
D	The set of m-dimensional points representing the users in U.
$\theta_{q,r}$	The Jaccard distance between q and r.
$RN_{\epsilon}(p)$	The neighbourhood of p in which the maximum distance. between any pair of points is ϵ.
G	A social group of users.
δ	The minimum number of joint visits.
$minPts$	A density threshold.

List of Algorithms

List of Figures

[1]Complementary Cumulative Distribution Function
[2]Maximum Likelihood Estimation

List of Tables

Chapter 1

Introduction

1.1 A Historical Context

The research into human presence and movement behaviour gained attention after the study by the geographer Finch in 1939 [41] in which he describes regions as knowledge objects. In the study, he summarises the unity of a region as forces and activities that link people to norms in the social world and objects in the real world - to learn the basis of a region geographers study static objects in order to understand dynamic processes; for instance, they study roads to learn about transportation within a given region. In contrast to Finch study [41], the traditional geographic research may have diverted the attention away from the dynamics aspects of the human presence and movement behaviour because material objects and their spacial distribution seem to enjoy a superior importance compared to the less favourable dynamics aspects [53]. As a result, the research into the human presence and movement behaviour seemed to have been forgotten until Ives and Messerli's study in 1981 [62] in which the two geographers recognised that making decisions in relation to hazardadjustment within an observed geographical area is critically dependent on both natural and dynamics aspects which include, amongst other factors, population movement. Although the aforementioned research may not have given the spacial aspect enough consideration but nonetheless provides good insights about the human presence and movement behaviour within a context of a unified geographical region.

1.2 Analytics of Human Presence and Movement Behaviour in The Age of Big Data

Nowadays, the precipitously increasing amounts of detailed information generated by wireless communication technologies such as GPS and Wi-Fi, provide an invaluable opportunity to study different aspects of presence and movement behaviours of people within a given environment such as an organisation office complex, a university campus or even a city. Moreover, the pervasiveness of these technologies increases peoples ability to access information, which undoubtedly influences the way the observed environment operates, and it is therefore essential that we develop the theoretical frameworks and the real-time monitoring systems in order to correctly understand how the presence and movement of people and its dynamics reshape the structures of such environments.

1.2.1 Urban Human Mobility

A thorough correct understanding of movement behaviours of people in urban environments can play a critical role in addressing challenges such as urban planning [108, 116], constructing smart systems for traffic forecasting [81], understanding crowd behaviour and event participation in mass gatherings [45], and developing effective epidemic control measures [2]. At an individual user level, a good knowledge of movement behaviours of individuals plays a crucial part in building smart mobile *recommendation* and *prediction* applications; namely, by capitalising on the pervasiveness of communication technologies such as GPS and cellular network location tracking which provide rich and detailed information about the locations that individual people visit. Employing such information, with all the rich knowledge that it contains about visited locations, is the foundation for building smart applications that address various urban challenges ranging from air pollution [115], traffic congestion [82] to finding a suitable restaurant [9, 79].

1.2.2 Human Presence Within a Specific Environment

The ubiquitous Wi-Fi infrastructure in many environments, such as universities and office buildings has made the Internet more accessible to a wide range of people in these environments, which consequently has given rise to new opportunities and challenges. For example, a university, which usually includes a variety of different spaces such as teaching rooms, laboratories, offices, retail, and residential buildings, can utilise access information to its Wi-fi network in order to learn about the realtime spatial occupancy of its facilities.

Gaining insight about spatial occupancy can be useful to a number of application areas including: resource allocation, surveillance, and the provision of basic facility services such as heating [8]. While a university can exploit its Wi-Fi network access information to learn about the usage of space within its campus [97], it can also capitalise on the such information to learn about the social groups that exist within its community, and how such groups interact with the available space [33].

1.3 Research Motivation

The proliferation of Wi-Fi and GPS enabled mobile devices and the vast amount of detailed information that these devices generate when accessing the Internet present an unmissable opportunity for studying the presence and movement of people within an observed environment. Motivated by such an opportunity, this thesis studies three main aspects of the human presence and movement behaviour: spatio-temporal movements, user identification, and social grouping. Contrary to previous research papers [4, 27, 68], it investigates all three aspects in order to develop a coherent view of the human presence and movement behaviour, addressing a set of specific challenges, which we briefly outlined hereafter.

1. **Prediction of Next Location of Visit**: Predicting the future location that an observed user will visit next is usually obtained by employing a *one-per-user-model*, i.e. a single user model, which exclusively comprises the historical mobility record of such a user [31]. Such a prediction approach has a number of limitations; for example, a model constructed exclusively from past mobility behaviour of a specific user would most likely perform poorly when utilised for predicting a future location that the observed user has never previously visited. This thesis attempts to find an alternative prediction approach that mitigates such a limitation.

2. **Identification of Mobile Users**: It has been established in a previous research paper [27] that only a small number of spatio-temporal points are enough to uniquely identify an individual user by utilising his/her mobility traces. This means, if a user u visited the set of locations $\{a, b, \ldots, z\}$ then only a small number of these locations would be enough to prove the uniqueness of the mobility traces of u. However, in this thesis we argue that a profile constructed and constrained to such a small set of spatio-temporal points - e.g. a profile that only includes information about 'home' and 'work' as two location points - would have limited benefit in the context of predicting and recommending locations to mobile users. Indeed in such contexts,

finding a distinct set of data that makes the individual unique is not the focal point. It is much more useful to have a rich profile that, in addition to being unique also reflects the individuals interests in terms of the places that s/he visits and the activities that s/he undertakes. Such a profile clearly offers a distinct advantage where it allows grouping together individuals with similar interests and tastes in terms of the locations that they visit. The ability to create such a grouping is the foundation upon which collaborative prediction and recommendation systems are developed [113, 117]. This thesis investigates the possibility of constructing a dynamic method of identification using mobility data which, for each individual user possess measurable variations that make it suitable for *"mobility fingerprinting"* [32].

3. **Detecting Class Attendance**: The growing number of new students and the courses offered at universities in recent years - due to competition between universities in attracting a larger share of new students [16, 56] - causes an increasing difficulty for campuses estate managements to correctly allocate the limited resources that are available to them. In order to mitigate such difficulties, universities are actively seeking new methods for estimating spatial occupancy. Consequently, there has been a growing interest in exploiting existing technologies, such as Wi-Fi, in order to track the human presence and movement behaviour on campus. Utilising existing Wi-Fi network in tracking attendance has a direct benefit, in saving costs, compared to other specialised tracking technologies which usually involve significant installation and running costs. In this thesis we investigate the detection of social groups that are formed as a result of attending learning activities at a university. We discuss how such detected social groups give insight about student attendance which can be utilised in estimating the real spatial occupancy in a learning environment.

4. **Characterisation of Space based on Visiting Behaviour of Mobile Users**: The numerous activities that take place within an observed learning environment such as a university campus determine, to a large extent, the kind of social interactions exhibited by the users in such environments. In this thesis, we attempt to understand the rules that govern these social interactions through analysing large collections of Wi-Fi activity traces of mobile users. More specifically, we are interested in whether we can characterise locations based on the visiting behaviours exhibited by social groups within such an environment.

1.4 Research Questions

This thesis considers two environments: a city environment represented by an average-size city in Europe and a learning environment represented by a university campus. For each of these environments we utilise a large collection of mobility traces: for the city environment we utilise a GPS data set from Nokia, and for the learning environment an Eduroam network log from Birkbeck, University of London, is utilised. We specifically focus on addressing three main research questions about the human presence and movement behaviour:

Q1. *Where and when do people move*? Utilising the spatio-temporal records of past movements, we are particularly interested in predicting the location that an observed user will be visiting next.

Q2. *How to uniquely identify people from their movement historical records*? The focus in addressing this question is not on whether a unique set of movements that characterise an observed user can be found. However, we are interested in building a profile made from a user's record of past movements that in addition to being unique can also be useful for the identification of such user from a short trail of recent movements.

Q3. *How do people interact*? We are specifically interested in social groups detection, particularly those groups that are linked to attendance of leaning activities at an observed learning environment; for example, a group of students that attend regular class sessions. Furthermore, we are interested in detecting those groups that visit locations such as a coffee-shop in order to socialise.

Q4. *How visiting behaviour characterises space*? We concentrate on the social groups' visiting behaviour exhibited at different locations within a learning environment. We particularly interested in those groups that are linked to the attendance of leaning activities, where we investigate the hypothesis that the distribution of a social group inter-visit duration, i.e. the waiting time between visits made by the same social group, follows a uniform distribution for location where *formal* activities, such as attending a meeting or a learning session, take place.

1.5 Contributions

Previous research works [4, 27, 68] have predominantly addressed two out of these three aforementioned questions, at most. Conversely, this thesis investigates all three questions in order to develop a coherent view of the human presence and movement behaviour. More specifically, we create stochastic models for movement prediction and user identification. We also devise a set of clustering models for the detection of the social groups within a given environment. Moreover, we propose a model for the characterisation of locations based on the social behaviour exhibited by mobile users when visiting these locations. The thesis makes the following contributions:

1. Proposes a family of predictive models that allows for inference of locations though a collaborative mechanism which does not require the profiling of individual users. These novel prediction models utilise suffix trees as their core underlying data structure, where predictions about a specific individual are computed over an aggregate model incorporating the collective record of observed behaviours of multiple users.

2. Defines a *mobility fingerprint* as a profile constructed from the users historical mobility traces. The proposed method for constructing such a profile is a principled and scalable implementation of a variable length Markov model based on n-grams. Furthermore, it demonstrates how the proposed fingerprinting method can be utilised in creating unique profiles for landmarks by successfully applying it to the *Next Location Prediction problem*.

3. Proposes novel density-based clustering methods that discover social groups by analysing activity traces of mobile users as they move around, from one location to another, within an observed environment.

4. Presents a novel model, which classifies locations into *formal* and *informal* locations on the basis of the visiting patterns exhibited by social groups detected at those locations.

1.6 Data Processing

The aforementioned research about the human presence and movement behaviour [4, 9, 27, 68, 79], could not have had a considerable impact without the large data sets that the researchers of these works had at their disposal. However, processing large volumes of complex data poses a serious challenge in terms of storage and performance [54]. In order

to overcome this challenge and exploit such kinds of complex data, numerous data mining and machine learning methods have been proposed [39]. One of the research areas that had a considerable share of these proposed methods is clustering, where several new algorithms have been devised [17, 35, 39, 83]. In this thesis clustering has been a central technique of some of our proposed methods, namely *Social-DBSCAN* and *Temporally-Restricted-Social-DBSCAN*, which we discuss in Chapter 6, and *SocialDBC*, which we describe in Chapter 7.

1.6.1 Data Utilised in This Thesis

We evaluate the proposed methods in this thesis on two data sets; namely, Nokia Mobile Data Challenge (Nokia MDC) data set and Birkbeck's Eduroam network log, which we outline hereafter.

1.6.1.1 Nokia Mobile Data Challenge (Nokia MDC) Data Set

In early 2009 Nokia Research Centre in Lausanne and its Swiss academic partners, namely Idiap[1] and EPFL[2], launched a campaign to create large-scale mobile data research resources [71]. Shortly thereafter, Nokia and its partners started the Lausanne Data Collection Campaign (LDCC[3]), an initiative to collect a longitudinal smart-phone data set from about 200 participants for over a year in the region of Lake Geneva. Nokia had an intention right from the start of the campaign to share the resources from this campaign with the research community, and thus launched the Mobile Data Challenge (MDC[4]), the challenge in which Nokia and its partners offered researchers an opportunity to study a data set that includes rich mobility, communication, and interaction information. The MDC had two research avenues: an Open Research Track and a Dedicated Research Track. Researchers who took part in the Open Track had the chance to propose their own tasks based on their research interests. On the other hand, researchers that took part in the Dedicated Track had the option of undertaking up to three different tasks to solve: prediction of mobility patterns, recognition of place categories, and estimation of demographic attributes. Experimental protocols and evaluation measures for assessing and ranking all contributions have been clearly defined for each of these tasks [71]. The LDCC data set was divided into four data parts for the benefit of the different tasks of the MDC:

[1] Idiap Research Institute
[2] École Polytechnique Fédérale de Lausanne
[3] Lausanne Data Collection Campaign
[4] Mobile Data Challenge

1. *Set A*: The shared training set for the Dedicated Trak tasks.
2. *Set B*: The test set for the demographic attribute and semantic place label prediction tasks.
3. *Set C*: The test set for the location prediction task.
4. *Open Set*: The set for all Open Track participants.

Out of those different sets of the MDC, this thesis only utilises the *Open Set* which, unlike the other MDC data sets, contains geo-location information (see sample data shown in Table 1.1).

User ID	Record Time	Time from GPS Satellite	Altitude	Longitude	Latitude	Speed	...
1234567	39363	53256	0.93599996567	6.63443099515	46.5128673917	63	...

Table 1.1: A sample Nokia MDC Open Set

1.6.1.2 Birkbeck's Eduroam Network Log

Birkbeck, University of London is one of the participant of Eduroam network [42, 105], a WLAN service developed for the international education and research community that gives secure, world-wide roaming access to the Internet. Birkbeck IT Services (Birkbeck ITS) provided us with a data set of Eduroam access information for the whole university for the period, from the 1st of October 2013 to 10th of April 2015. This portion of the data set used in this thesis comprises 223 locations and 204.6K users, who come from 2462 institutions and departments. The 223 locations given in this data set are divided between 11 of the 17 sites of Birkbeck's Bloomsbury campus in central London. User ID, access point location, connect time, duration of session, MAC address of user's device and affiliation email address are the basic information for each processed record (see sample data shown in Table 1.2).

User ID	MAC Address	Connect Time	Disconnect Time	Session Duration	AP Location	...
1234567	00:03:ff:60:fb:fn	12/06/2013 10:40	12/06/2013 11:20	40 min	MaletSt-319	...

Table 1.2: A sample Eduroam network log

A comparison of the two types of data sets utilised in this thesis can be found in Table 1.3.

A summary of a method used to process the data in this thesis is provided hereafter.

Data set	Eduroam log from Birkbeck	Nokia MDC
Data type	Wireless network traces	GPS data
Spatial Resolution	meters	meters
Scale of Area Covered	Campus or work location	An area covering a city region
User's speed	No	Yes
Path between two locations	Approximate	Exact
Number of users	204.6K	38

Table 1.3: A comparison between the basic features describing the users mobility in the two data sets used in this thesis

1. Raw data collected from Wi-Fi access points which are positioned in widespread locations across the university campus. Each user's device generating the data is identified by its MAC[5] address and each data point also records a time-stamp (see sample data shown in Table 1.2).

2. Each router provides meta-data that can be employed for identifying the semantics of the location in which the router is situated; for example, a router may be located inside a classroom or in an area in a coffee-shop, or a cinema.

3. The raw data is transformed, by applying clustering and other data analysis methods, into trail data which is stored in suffix tree data structures or a DBMS[6]. A suffix tree data structure captures the sequences of movements embedded in the trails which the users followed based on the time-stamps provided in the raw data. It can also be queried for a specific sequence of movements, hence it can be employed for individual users' tracking.

1.6.1.3 A Brief Note about Data Privacy

All sensitive data items, of the data sets utilised in this thesis, such as the user's email/name and their device's MAC address, have been anonymised to allow the type of analytics provided in this thesis to be carried out without compromising the user's privacy. Moreover, we do not attempt to use location in a way that compromises privacy, e.g. by displaying actual locations on maps. These data processing and related security and data management provisions have been approved by Birkbeck's research ethics committee, which ensures strict compliance with EU law on data protection and privacy (GDPR[7]).

[5]Media Access Control Address
[6]Database Management System
[7]General Data Protection Regulation

1.7 Outline of the Thesis

With exception to this introductory chapter the thesis is briefly outlined hereafter.

Chapter 2. We provide a critical review of works carried out by other researchers.

This thesis then divides into two main parts. The *first part* investigates the spatio-temporal movement where we predict the future locations of visit based on when and where people had been in the past. We also investigate the identification of users from their historical movements. We propose stochastic models for movement prediction and user identification which we evaluated on the Nokia MDC data set (see § 1.6.1.1).

Chapter 3. We study the *collective* model and the one-model-per-user approaches in the context of the next location prediction problem - the problem of predicting a user's subsequent location of visit, taking into consideration the time and location information of where the user had been in the past. Furthermore, we study the effect of the length of the user record of the most recent temporal locality used in the prediction of the next location of visit, and assess the relative loss of accuracy when smaller data records are provided so as to establish the exact *trade-off* involved. We evaluate our performance of the proposed prediction models, i.e. the single user model and the collective multi-user model, by using MAE and RMSE error metrics. We show how to use these two metrics to determine the number of suggested (the top-k) locations which are most likely to include the observed user's correct next location of visit. Moreover, we study the merits of HM Score in assessing the accuracy of the proposed models.

Chapter 4. We investigate whether the trails generated from users' mobility traces have sufficient measurable variations which allow for fingerprinting of movements of those users to whom these traces belong, i.e. can we verify and measure the uniqueness of individual users movements. Also assuming that the users have different *mobility fingerprints*, we examines the *identifiability* of the correct user from an observed mobility trail, i.e. whether a user can be identified from his or her trail of movements. The same chapter investigates whether the size of the fingerprint can be reduced while retaining *identifiability*, and to this end, it attempts to find a minimal fingerprint that can be employed to correctly identify an observed user from a short record of movements. It also investigates whether the proposed fingerprinting method can be extended to create unique profiles for landmarks

and whether such fingerprints can be used for location prediction.

The *second part* of the thesis considers the social grouping concept (how do people interact). The clustering models proposed for the detection of social groups, and location classification, within an observed learning environment, have been applied to the Eduroam data obtained from Birkbeck, University of London (see Subsection 1.6.1.2).

Chapter 5. We present a comprehensive analysis about the human presence within a university campus where it provides a thorough analysis with respect to the four types of patterns contained in the data: the social, the spatial, the temporal and the semantic patterns, giving an insight into how people presence shapes the dynamic structure of such an environment. For each of these types of pattern: the social, the spatial, the temporal and the semantic, the chapter defines a list of metrics, which we utilise to interpret the observed behaviour captured in the data.

Chapter 6. We introduce social density-based clustering methods that use WLAN traces in order to detect granular social groups of mobile users within a university campus. The proposed clustering methods rely on the underpinning semantic context for parameterisation, i.e. utilising information from the semantic context to determine the values of the clustering algorithm parameters. The same chapter also estimates the actual level of attendance of learning activities - linking the discovered social group that regularly visits an observed location and the learning activity that takes place within the same context allows us to estimate the attendance level of a targeted learning activity.

Chapter 7. Herein we propose a density-based clustering method that discovers social groups by utilising activity traces of mobile users. We detect the social groups on the basis of the activities taking place at observed locations within a university campus. It also proposes a framework for inferring the type of an observed location, by using the patterns of visit extracted from Wi-Fi activity traces.

Chapter 8. We provide our concluding remarks and a summary of future work.

Chapter 2

Critical Review

2.1 Overview

This chapter provides a critical appraisal of the research papers concerning the analytics of human presence and movement behaviour that are available in the technical literature. Numerous research investigated the possibility of exploiting activity traces of wireless communications in order to gain insight into the human presence and movement behaviours within a given environment. We review some of these papers in relation to the four data aspects: the social, the spatial, the temporal and the semantic aspects. We are particularly interested in the prediction of the user's next location of visit by using GPS[1] data, the identification of the user from his or her trails of visited locations, and the detection of social groups that the user maybe associated with. The discussion in this chapter is organised as follows: in Section 2.2, we review the research efforts in addressing the next location prediction problem, and in Section 2.3 we discuss the methods utilised for the identification of mobile users. In Section 2.4 we provide a critique of the papers relating to the detection of social groups by employing Wi-Fi activity traces, and conclude with a thorough discussion in Section 2.5.

2.2 Prediction of Next Location of Visit by using GPS Data

With the mobile phone becoming widespread human mobility data is now captured and stored, as never before. This motivated the research into human mobility patterns which in recent years started to receive a lot of attention as increasingly more volumes of detailed

[1]Global Positioning System

mobility data become available. The advancement and pervasiveness of wireless communication technologies did not only cause an increase in the number of users taking part in human mobility studies but also meant that the areas considered in such studies are much larger in size than ever before. As a result, we have interesting findings from some of the recent research about users' mobility patterns. It has been established that there is high regularity in mobility patterns exhibited by individuals despite the size differences of the areas in which people move [49, 93]; for instance, users in a city environment oscillate between home and work every weekday while students at a university campus regularly visit a set of specific rooms to attend classes [33]. Setting aside any strange or unusual visiting habits, the researchers in [92] were able to find universal laws that govern the users mobility behaviour when visiting new places or revisiting locations that they have already been to in the past. In [28] researchers found that members of the same social group exhibit the same mobility behaviours. A similar finding was provided in [33], where users from the same social group are likely to visit the same location when they are in the company of one another. A key benefit that can be drawn from these findings is that a reliable model can be developed for inferring users' future movement. In the remainder of this section, we provide a critical review of selected works, from the technical literature, that address the problem of predicting the next Location of visit by using GPS Data.

2.2.1 Extracting Locations of Visit From Raw Data

An observed user's *location of visit* can be a place that a user frequently visited in the past or a place that s/he stayed at for some significant time. Such a location does not have to be a place that the user visits in order to socialise with other people; for instance a restaurant. It can be any frequently visited place such as a petrol station or a busy junction in the user's daily journey to work. Figure 2.1 highlights the set of locations learned from one individuals GPS recordings obtained from the Nokia MDC data set [71]. Some of these locations shown in the figure, i.e. Figure 2.1, correspond to geographical meaningful locations such as "home" or "work place" but equally there are other locations that do not correspond to such meaningful geographic places; for instance a busy junction on the road. Generally, a mobility trace in a GPS data set is a sequence of latitude and longitude pairs where each pair is associated with a time-stamp. In order to extract the locations of visit from such data a host of methods have been proposed over the past few years [5, 18, 31, 84, 118]. For example, in [5] a dual step method was proposed for extracting significant locations of visit which are later analysed to predict the next location of visit using a Markov model [84]. In step one, the significant locations of visit

are detected by using the points where the mobile device loses connection to the GPS satellites. In step two, clusters of locations are formed by using a variant of the K-Means algorithm. At the start of the clustering process, the locations clusters are centred at K selected points with a given radius - a cluster with a large radius here may correspond to a city while a cluster with a small one may correspond to a campus or an office building. The drawback of this method comes from its dependence on the lose of signal in order to detect locations of visit, i.e. the method would fail to detect locations that have continuous reception of signal; for example, it would fail in detecting open locations such as an open market with stalls for selling goods where the signal reception is likely to be uninterrupted. On the other hand, the method would probably succeed in detecting office buildings and other similar locations which are likely to have no GPS signal reception. The authors of [118] proposed a clustering method called DJ-Cluster which uses *density* and *joining* concepts in order to extract significant locations of visit. Similar to other density-based methods, a dense point in this method is a point that has a total number of neighbours greater than or equal to a user-defined minimum threshold required for all dense points. Clusters are then created by *joining* density points together in the same cluster if they have common neighbouring points between them. An improved method, proposed by the same authors, removes a GPS reading if it has speed greater than zero or if its distance from the previous reading is below a given threshold. The tests result of their new method indicate an improvement over the K-Means in terms of precision and recall, and DBSCAN [35] in terms of time and memory requirements. In [18] a semantics-enhanced clustering algorithm, called *SEM-CLS*, was proposed for extracting semantically meaningful locations. This method separates semantically different locations into different clusters and merges those locations with similar semantics into the same clusters.

2.2.2 Next Place Prediction Models

To decipher the complexity of predicting human mobility, many approaches have been proposed for building models that can accurately predict individuals' future locations of visit. Generally, these approaches can be divided into three major categories based on the perspective from which the data is being considered: spatial, temporal, and joint spatio-temporal approaches. Researchers have investigated the user's spatial patterns on mobile data and various prediction approaches have been proposed, such as [114]. Other proposed methods that rely on the user's temporal patterns in order to predict the next place of visit, as shown in [3]. However, discovering the correct temporal patterns in human mobility is challenging, since temporal behaviour includes much more uncertainty

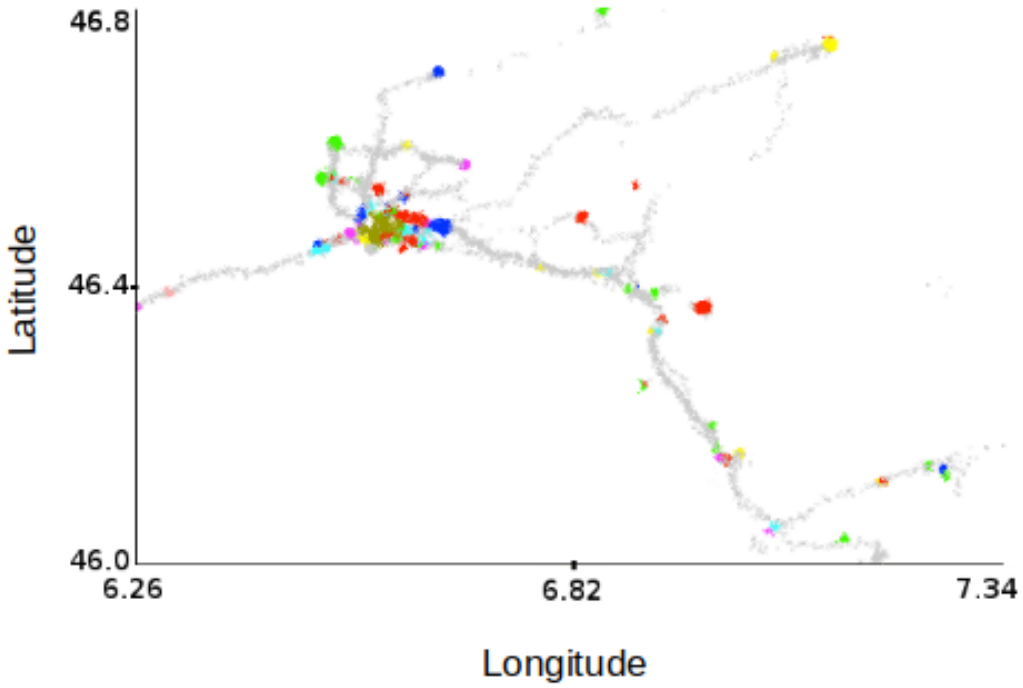


Figure 2.1: The locations visited by one of the users as detected by the DBSCAN algorithm [36]. The grey colour shows clusters of GPS points identified as noise while the other colours show the discovered locations of interest.

in comparison to the spatial behaviour [94].

In [87], Scellato et al. proposed a spatio-temporal framework, called NextPlace, which used non-linear time series analysis of users' arrival time and residence time to predict temporal behaviour. Chon et al. [23], used fine-grained and continuous mobility data to evaluate several mobility models. They argued that the granularity of mobility data used in the literature is too coarse to precisely capture users' daily movement patterns. Although joint Wi-Fi/Bluetooth traces were used as opposed to GPS data, Vu et al. in [103] introduced a framework for building predictive models of people movement. The proposed framework used a type-of-day categorisation (such as weekday and weekend) to filter redundant information from users' historical data. Noulas et al. on the other hand, studied the problem of predicting the next venue that a mobile user will visit, by extracting features from check-ins data of Foursquare users. The extracted features exploit information about transitions between types of places, movement between different venues, and spatio-temporal patterns of user check-ins [79]. They proposed two learning models, based on linear regression and M5 model trees, which combine all individual features.

Using a list of thousands of candidate venues, the proposed supervised methodology which combines multiple features offered high levels of prediction accuracy, where M5 model trees was able to rank in the top fifty venues one in two user check-ins.

2.2.2.1 A Single-user Model Versus a Multi-user Model

Prediction models of future locations of visit have been predominantly implemented using a one-model-per-user approach. For example, Krumm [68], used a Markov model for making short-term route predictions for vehicle drivers. Ashbrook and Starner [4] suggested a model in which locations are incorporated into a Markov model that can be consulted for use with a variety of applications in both single-user and collaborative scenario where multiple single-user models can be shared. Unfortunately, it is not clear how they evaluated their models apart from showing that the predictions for their single user model were compared against "random chance". Also they did not address the situations in which the user has no mobility history to be exploited when predicting future location of visit. Moreover, sharing multiple single-user models inevitably raises concerns relating to the privacy of users' information being compromised; for example, by a service provider gaining access to a user's mobility history embedded in a single-user model for such a user. Contrary to the modelling style adopted in [4] and [68], Chapter 3 of this thesis presents a collective (i.e. a multi-user) next location prediction model which does not specifically store an identifiable individual user mobility records in order to predict future location of visits for such a user. This collective model is a principled and scalable implementation of a variable length Markov model. Furthermore, the same chapter, i.e. Chapter 3, presents various models that address the situations in which the user has no mobility history to be exploited for inferring future locations of visit.

2.3 Identification of Users Through Mobility Traces

2.3.1 Background

Technological device *fingerprinting* relies on measuring the small differences present in each device which makes it distinguishable from the other devices of its type. It has been long established that devices such as Cameras as described in [22, 73] and typewriters as in [58] can be distinguished from other similar devices through *fingerprinting*. In [29], Peter Eckersley investigated the degree to which modern web browsers are subject to such fingerprinting by analysing the information sent to websites upon request. By introducing the concept of fingerprinting to distinguish between web browsers Eckersley has thus set

the scene for the identification of individual users through data extracted from their web browsing activities. Indeed, an individual user is identified through their browser history, i.e. the list of URLs they have been browsing which are surely unique to them, just as much as their biological fingerprint is [55]. In the same article, i.e. [55], Brian Hayes explains how we now also have what he refers to as "data identity", defined by various combinations of traits that distinguish us from anybody else on the planet. This idea about *data identity* was well supported by the work carried out by Sweeney of Harvard University [98] in which she showed how, by using only a small set of simple demographic information such as the date of birth, the zip code, and the gender, we can identify an individual from the rest of the population. Furthermore, the authors of [112] describe a system called WiFi-ID which extracts unique features that capture the walking style of a person, and thus allow for the unique identification of such an individual, by analysing *the channel state* information.

2.3.2 Uniqueness of Mobility Traces

In [27], Yves-Alexandre et al. proposed a formula that determines the uniqueness of individual mobility traces. A key result of their work is that they showed that the uniqueness of human mobility traces is high and that individual users mobility data are likely to be identified using information about only a few outside locations. In the same research work, i.e. [27], Yves-Alexandre et al. further showed that only four spatio-temporal points are enough to uniquely identify 95% of the users in the large data set that they used for evaluation. This means, if a user u visited the set of locations $\{a,b,\ldots,z\}$ then only four of these locations would be enough to prove the uniqueness of the mobility traces of u. This is very much consistent with our finding presented in [32] and in Chapter 4 which provides a detailed discussion about the uniqueness of the individual users' mobility fingerprints. However, our work differs substantially, because in addition to creating a unique user profile, we can also employ such a profile to identify the user from a short record of observed movements; for example, if $\{e,f,g,h\}$ denotes some observed mobility trail, then we can employ the fingerprint constructed from the user's historical record of visit (e.g. to the locations $\{a,b,\ldots,z\}$) to correctly predict that the observed trail was created by the user u. A substantial part of Chapter 4 of this thesis is dedicated for the investigation of the relationship between the user *identifiability* and the fingerprint uniqueness as well as the implications when the fingerprint is compressed.

2.4 Detection of Mobile Users Social Grouping by using Wi-Fi Activity Traces

Studies involving Wi-Fi networks data analysis can be divided into two broad categories: descriptive versus predictive analysis. While descriptive research on characterisation of user mobility in Wi-Fi networks explore various features such as the time duration the user spends connecting to an Access Point (AP[2]), and the amount of data a user sends and receives over the network, predictive studies can be classified according to the modelling approach adopted in such studies. Amongst the common modelling methods utilised in previous studies are: clustering [65, 103], Support Vector Machines (SVM[3]s) [74] and Markov models [69, 72, 107]. In this section we review both the predictive and the descriptive research works that are available in the literature focusing specifically on the social dimension of the human presence within an academic institution.

2.4.1 Social Groups of Mobile Users

Using data collected from a hundred mobile phones over a period of nine months, the authors of [28], proposed a system for complex social systems' sensing. They were able to detect social patterns in daily user activity, infer user relationships, discover socially significant locations, and thus model the rhythms of observed organizations by using standard bluetooth-enabled mobile phones. Static bluetooth device IDs were used as an additional indicator of location, and this was shown to provide a significant improvement in user localization, especially within indoor environments such as an office building. The authors of [59] proposed a method for extracting interaction patterns and social behaviour of mobile users by using passive WiFi monitoring of *probe requests* and *null data frames* that are sent by smart-phones. They are able to discover proximity relationships, occupancy patterns, and social interactions among users by analysing the temporal and spatial correlations of the Receive Signal Strength Indicators (RSSI[4]) of packets from these low rate transmissions. Although results of conducted tests, which used commodity off-the-shelf smart-phones and WiFi Access Points, demonstrate that the proposed method is capable of detecting social relationships and interactions in a non-intrusive manner, the study was conducted on a very limited scale. In [30] and in Chapter 6 of this thesis, namely in Section 6.5, we discuss a method for detecting classroom friends by using a data set representing a full snapshot of Wi-Fi usage of a whole university for a period covering a

[2]Access Point
[3]Support Vector Machine
[4]Receive Signal Strength Indicators

full academic term.

2.4.1.1 Attendance of Learning Activities

In [97] an occupancy sensing system for a real university campus environment was proposed. The researchers conducted a lab experiment in order to evaluate various commercial sensors in terms of cost, ease of operation, and accuracy. Deploying beam-counter based system in 9 real classrooms of varying sizes across their university campus, they collected data over a period of 12 weeks covering more than 250 courses. Employing detected course attendance patterns and classroom occupancy, they developed an *off-line* method that dynamically allocates courses to classrooms, and thus they managed to make gains of over 50% in room related costs. In [76] the authors explored the use of Wi-Fi for estimating attendance in a dense university campus environment. They proposed new methods for distinguishing and filtering out WiFi-connected users outside an observed lecture room, and feed such data to a regression model in order to estimate room occupancy. The authors of [85] analysed data from a Wi-Fi network at technical university using different granularities (each individual access point, groups of access points, entire network) in order to study the network usage. Their work investigated whether students attending a lecture use the wireless network differently in comparison to the way students not attending a lecture do. By employing a supervised learning approach based on Quadratic Discriminant Analysis (QDA[5]) they are able to classify rooms into empty and occupied spaces. Although the proposed method can detect room occupancy, i.e. rooms being empty or occupied, it falls short in detecting attendance of lectures as it has no means of tracking individual student's class attendance. In Chapter 6 of this thesis, namely in Section 6.4, we discuss a method for estimating class attendance by tracking the attendance of individual students over the course of a given academic term of 11 weeks. In [119] the researchers attempted to measure students' behaviour in classroom-based courses in a large-scale study. They proposed a system, called EDUM (EDUcation Measurement) to characterise educational behaviour at a large university campus. They investigated a number of behaviours including class attendance, and late arrival to lectures as well as early departure. Their research work had some interesting findings; for example, they detected class attendance and what time of day it reaches its highest and lowest levels, the most hard-working day of the week by using measures such as the *attendance ratio* and the *late arrival ratio*. While their proposed method employs data from multiple sources including Wi-Fi data, in Chapter 6 of this thesis, we discuss how we detected class attendance by inferring session attendance

[5]Quadratic Discriminant Analysis

utilising patterns extracted only from Wi-Fi activity traces. Moreover, the ability to filter *noise*, i.e. bystanders (individuals who are not part of the intended class but nonetheless appear to be part of it), is a key factor in developing a successful method that can detect the attendance of an observed class. In the same chapter, namely in Section 6.4.3, we discuss two methods for noise removal: *Noise Reduction* and *Attendance Coherence*. In [119] which employs data from multiple sources, the removal of *noise* merely depends on how far a connected mobile device is located from the Access Point.

2.4.2 Spatial Classification

Unfortunately, the research in *space-based modelling* (i.e. models that focus on space) of the human presence and movement behaviour falls short in devising laws that describe space patterns [109]; for example, how the interactions and occurrences of activities are timed in spatial distribution. Modelling space from the perspective of time allows for the spatial organizations and temporal ordering of spatial functions [109]. Due to the lack of research contribution, we do not have a good theoretical understanding of this area [109]. However, in Chapter 7 and in [33], we investigate the hypothesis that the distribution of a social group inter-visit duration, i.e. the waiting time between visits made by the same social group, approximately follows a uniform distribution for locations where *formal* activities, such as attending a meeting or a learning session, take place. We developed a model that learns a spatial classification in which the type of an observed location is predicted based on the patterns of inter-visits durations of detected social groups. The details of this model is discussed in great detail in Chapter 7 of this thesis.

2.5 Discussion

In this chapter, an overview of the major techniques in: 1) the prediction of the user's next location of visit by using GPS data, 2) the identification of a user from short trail of movement activity, and 3) the detection of social groups that the user may be associated with, was presented. In this section, we discuss key aspects that have not been considered in previous research in relation to these three areas.

2.5.1 A Collective Inference Approach

Many previously proposed location prediction approaches directly concentrate on inferring the next location without investigating whether, or not, the data being used might have latent clustering of mobility patterns such as "weekday" and "weekend". Discovering such

a clustering in the data from the outset may lead to a significant improvement in the prediction accuracy. Our proposed collective model, which is based on a single aggregate model incorporating the collective record of observed behaviours of multiple users, can exploit such an underlying clustering of the users to improve prediction accuracy. An exploratory investigation we carried out in this thesis revealed that there a significant variance in the average number and the kind of places visited by the users during the weekend period compared to that of the weekdays period. Capitalising on such a finding, we build separate models for each class of data as discussed in Chapter 3 of this thesis.

2.5.2 Assessing the Accuracy of Next Place Predictions

The *Hit and Miss Score* (HM[6]) which, also known as the "Hit Ratio" or the "Success Ratio", has been widely used to measure the prediction accuracy in domains such as Web usage mining [14] and in user mobility prediction [37]. It is normally computed as the number of successes divided by the total number of attempts made. However, if an observed model produces a set of predictions as opposed to a single one, the HM is used to measure the proportion of times the correct item has been included in the set of predicted items [14]. In the context of predicting users' mobility, our opinion, which is strongly supported by our tests' results shown in Chapter 3, is that HM on its own does not provide a sufficient assessment for the prediction accuracy, and thus employing additional metrics using the mean error, i.e. the *Mean Absolute Error* (MAE[7]) and the *Root Mean Square Error* (RMSE[8]), may be preferable.

2.5.3 Mobility Fingerprinting

Although there are various kind of reasons that motivate the correct identification of individuals, our motivation, in this thesis, stems from the desire to have a unique profile that encapsulates the individual's interests in terms of the places that they visit and the activities that they undertake. Therefore, we emphasise that the underlining purpose of our proposed *mobility fingerprinting* method, which we discuss in Chapter 4 of this thesis, is to provide a platform for constructing more robust context-aware mobile prediction systems that can equally be employed for user identification.

[6]Hit and Miss Score
[7]Mean Absolute Error
[8]Root Mean Square Error

2.5.4 Social Groups Detection

The ability to measure the proximity between co-located individuals, during a visit to an observed location, is a key factor in accurately inferring whether the individuals are socialising or visiting the target location for different reasons; for example, when two students visit the Coffee-shop after the class, but sit at separate tables. Unfortunately, the Wi-Fi data set we utilised for the evaluation of the social groups detection methods proposed in this thesis does not contain any proximity information. In [102], a novel system that enables a single Wi-Fi *access point* to localise devices within a distance of tens of centimetres was proposed. With such a system in place it is feasible to have rich data sets that contain information about the proximity between users visiting an observed location. Evaluating our proposed methods using such a richer data set will most likely increase the accuracy of the obtained results. Setting aside the lack of proximity information, in Chapter 6 of this thesis, we proposed a method for detecting class-room friends by detecting attendance of learning activities then detect social groups that visit locations such as the coffee-shop during break-times.

Chapter 3

A Collective Prediction Model

3.1 Overview

Models designed for predicting the location that an observed user will visit at a future time, are typically implemented using a one-model-per-user approach which cannot be employed for inferring collective or social behaviours involving other individuals. In this chapter, we propose an alternative that allows for inference through a collaborative mechanism which is cheaper to maintain and does not require the profiling of individual users. Specifically, we introduce a family of prediction models that utilise suffix trees as their core underlying data structure, where predictions about a specific individual are computed over an aggregate model incorporating the collective record of observed behaviours of multiple users. We evaluate the performance of these models on the Nokia Mobile Data Collection Campaign data set and find that the collective approach performs well in comparison to individual user models. We also find that the commonly used Hit and Miss score (HM) on its own does not provide sufficient indication of prediction accuracy, and that employing additional metrics using the mean error, i.e. *Mean Absolute Error* (MAE) and *Root Mean Squared Error* (RMSE), may be preferable.

3.2 Introduction

Predicting the behaviour of individual mobile users is a key factor for context-aware adaptation in mobile applications and systems, and moreover avails the foundation upon which mobile recommendation systems are developed. One approach to build a mobile recommendation system is to employ the one-model-per-user paradigm to predict subsequent locations that the user will visit [38]. However, by setting the goal "to build a user-specific

model that learns from his/her mobility history, and then apply the model to predict where the user will go next", one presents a tightly defined objective: this model must be constructed solely from personal data recorded specifically as the result of the behaviour of the particular observed user. Furthermore, the one-model-per-user approach has several further potential limitations:

1. In practice, most such models typically involve machine-learning techniques which can produce reliable inferences but only about behaviours that have been previously observed and incorporated into the model. However, they have relatively poor performance in situations when novel behaviours occur for the first time.

2. Focusing on the individual, such models cannot be employed for predicting collective/social dynamics, which are often the cause of interesting, and sometimes surprising, individual behaviour such as those resulting from cascading behaviour within social networks [6].

3. Individual user models are often sparse and do not contain enough information to make reliable predictions.

An alternative approach to the one-model-per-user is to predict the behaviour of an observed user by employing a single aggregate model incorporating the collective record of observed behaviours of multiple users.

In the next section, i.e. Section 3.3, we define the research problem and in the following section, we present our model starting with the description of mobility trails, which are the building blocks of our proposed model, by showing how they are constructed from users' mobility traces. In the same section, we present the suffix tree data structure and how it relates to the one-model-per-user and collective model approaches. We propose two location-independent prediction models in the "Temporal Models" section. We describe the data, experiments and the metrics used for assessing the prediction accuracy as well as providing an extensive evaluation of the experimental results, in the "Evaluation" section, i.e. Section 3.6. The concluding section in this chapter compares the one-model-per-user and the collective model. It also debates the merits of using *Mean Absolute Error* (MAE) and *Root Mean Squared Error* (RMSE) versus the *Hit and Miss Score* (HM).

3.3 Problem Definition

Within the scope of this chapter, we:

1. Investigate the limitations of the collective model and the one-model-per-user approaches in the context of the next location prediction problem - the problem of predicting a user's subsequent location of visit, taking into consideration the time and location information of where the user had been in the past.

2. We examine the effect of the length of the user record of the most recent temporal locality used to make inferences, and the relative loss of accuracy when reduced data samples are provided so as to establish the exact trade-off involved.

3. We evaluate our proposed approach using MAE and RMSE error metrics. We show how to use these two metrics to determine the number of suggested (the top-k) locations which are most likely to include the observed user's correct next location of visit. We also investigate the merits of HM in evaluating the accuracy of the proposed location prediction algorithms.

3.3.1 Contributions

Supported by empirical evidence, this chapter makes the following two main contributions:

1. It presents a collective approach for predicting the mobility behaviours of users. This approach is a principled and scalable implementation of a variable length Markov model [12] which allows for collaborative inference while it does not require the profiling of individual user behaviour.

2. It compares the prediction accuracy of the collective and the one-model-per-user approaches. It also examines the strengths and limitations of each approach.

3. It describes a more comprehensive approach, than previously proposed, to evaluating the mobility model's prediction accuracy.

3.3.2 Methodology

We have followed a specific methodology to construct the *collective model*, the details of which will be discussed later in the next Section 3.4. In brief, we construct a variable length Markov chain model [12], which we store in a suffix tree data structure that allows for flexible and high-performance querying. This representation is supplemented by specific

weighted spatio-temporal metrics of significance to estimate and rank the probability of computed inferences.

3.4 Modelling with Suffix Trees

3.4.1 Mobility Trails

The cornerstone of our approach is the use of *trails* as the principal data processing primitive for analysis and prediction. Our choice of trails was not arbitrary. For centuries, and in many ways, trails have been used as the basis for coordination between humans. For example, navigation trails provide route information and record information about paths to specific destinations. Aggregating multiple trails acquired over time across a particular environment is the technique humans often use to develop complete maps of a particular landscape, and subsequently assist navigation, especially in the context of exploration [48].

We consider a mobility trail of an observed user as the sequence of recordings, of the temporal and spatial information, of all the visits that the user makes in a day. Trails contain users' mobility patterns and they can be used for the provision of different services, spatial, temporal and social analysis [25]. However, a traditional drawback of trail analysis is that it requires considerable storage and computational resources to discover such patterns. To overcome this, we employ a trail-based analysis approach, which utilises suffix trees as the data structure for efficient storage, filtering and retrieval [10].

We view a user's mobility history as a directed graph, where vertices denote locations which the user visited and edges denote paths between such locations. Two locations are said to be connected if they have been visited in sequence by the observed user. In such context, a trail can be defined as a sequence of connected locations, such that the connections between locations are always directed. The connections in the trail are weighted with different usage meta-data such as the time-stamp indicating the time of visit of the destination location.

3.4.2 Detecting a User's Mobility Patterns

In contrast to using known landmarks which describe the positions of significant entities within the landscape that the users interact with, we concentrate on utilising the sequence of geographic coordinates recordings of the user's exact location. The basic assumption here is that over a period of time and as the user, in their daily routine, moves from one

location to another, some of the user's mobility patterns would have been captured in the aggregate of these coordinates recordings. To detect these patterns we apply the following procedure:

1. For every user, apply the DBSCAN[1] algorithm [36] to cluster the GPS data to identify the set of locations which are likely to be part of a daily pattern as opposed to just noise (Locations that have a number of visits below a certain threshold are considered noise and hence ignored - see Figure 2.1).
2. Compute the *centre of mass* of the locations in each discovered cluster.
3. Using a latitude/longitude grid, for each cluster, identify the grid cell(s) containing the centre of mass of the clusters.
4. Divide the day into equal time-units. (We choose 20 minutes as the basic time unit).
5. Compute the duration of visit for each cluster using the time of visit associated with the GPS readings. Then identify the time-unit(s) corresponding to each visit.
6. Construct the daily trail using the grid cells and the time-units computed in step 5.

The result of the processing given above is to obtain sequences of tuples where each tuple contains the following information:

⟨***user, time, day, location, meta-data***⟩

where

i) ***user*** : the ID of the particular person involved.

ii) ***time*** : the corresponding time-unit in which the GPS reading was recorded.

iii) ***day*** : the corresponding day of the week in which the GPS reading was recorded.

iv) ***location*** : the ID of a specific grid cell, which contains the centre of mass of the cluster of the visited location.

v) **meta-data** : a list which may contain information such as the duration of interaction, the exact time and date of visit of the destination location.

[1]Density-based Spatial Clustering of Applications with Noise

3.4.3 Suffix Trees

Trail analysis has a major drawback due to the fact that it may require considerable storage and computational resources to discover hidden patterns. To efficiently store the trails and their related meta-data we use a probabilistic suffix tree data structure [10] enhanced with meta-data needed to encapsulate different information and metrics. Our choice of this data structure, i.e. suffix trees, was motivated by the fact that suffix trees can maintain all captured information in a compact format, while being able to respond to queries in linear time in the size of the trail, and, in addition, being capable of responding to requests about any number of possible times, space and semantics related criteria. Suffix trees have been successfully employed in a number of domains such as anti-spam filtering [80] and computational biology, where they were used to address problems such as string matching applied to DNA[2] sequences [10].

3.4.4 Tree Representation

For our suffix tree representation, we opted for a design in which the nodes are labelled as opposed to the edges [80]. Also, due to the nature of the task undertaken in this work, our suffix tree uses a terminal character to determine the depth of the tree which, depending on the number of users in the data set, can grow very large in size. Furthermore, our trees are not limited in size which is a key factor that gives the model the ability to learn as more locations are being explored by the observed users. An example of this data structure is shown in Figure 3.1. For a thorough description, along with algorithms to efficient memory usage and improved processing speed, the reader is referred to [51].

3.4.5 The One-model-per-user

The one-per-user suffix tree model is based entirely on a single user's past mobility data. The history data is divided into trails representing the daily sequences of visits made by the observed user. For each such sequence of daily visits, the time and location information of each visit is encoded in a string object and the objects, in turn, are grouped together to form a trail. In order to make a prediction, the tree is presented with a *search-trail*, which is the trail representing the sequence of the most recent visits that the user made including the current visit. If a matching sequence of visited locations is found, the tree responds with a list of candidate locations. Each candidate location has been visited by the user, in

[2]Deoxyribonucleic Acid

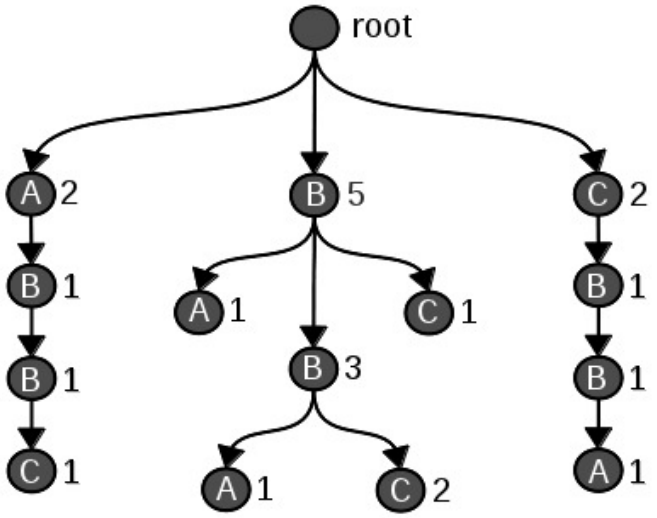


Figure 3.1: A suffix tree for the trails represented by the strings "ABBC", "CBBA" and "BBC". The letters denote the individual visits made to the locations in each trail and the numbers show the frequency of visit to each location.

the past, immediately after making the sequence of visits given in the search-trail. Those *top-k* candidate locations with the highest frequency of visit are predicted as the next k locations, i.e. each of these k locations is most likely to be visited by the user immediately from his/her current location.

3.4.5.1 Predicting the Next Location with Suffix Trees

Suppose that we have a training set A and a test set B of mobility trails of an observed user u. Let S be a suffix tree built from the trails in the set A. Suppose that we have a trail, T $= t_1$, t_2 , . . . , t_n obtained from our test set B, where t_i , $1 \leq i \leq n$, represent the ordered locations visited by the observed user u, and we wish to discover the location t_{n+1} that the user u is likely to visit next. We can apply *Algorithm 3.1*, called Predict, based on [63], to predict the next location:

3.4.6 The Collective Model

The collective suffix tree model is a joint model over the population of all users. It enables prediction of the next location of visit based on the past mobility data of multiple users. The training data for the model is made of the union of all the training sets used for building the individual per-user suffix tree models. The data identifying specific users (i.e. userID) is completely excluded before building the model. The location data is transformed

```
1: Predict(T, S, k)
2: let s be the longest suffix of T in S
3: if s is empty then
4:     return the top-k popular/time spent locations;
5:     # these will be children of the root
6: else
7:     return a ranked list of the k most popular
8:     locations directly reachable from s;
9:     # these will be children of the last location in
10:    # s, where s is a path from the root
11: end if
```

Algorithm 3.1: Predict

into grid-cell-ID which captures most of the mobility information but excludes the actual geographic coordinates.

3.5 Temporal Models

The motivation behind the simple models described in this section is to address the lack of matching historical behaviour which the suffix tree requires when predicting the observed user's future behaviour. Both of the two models: *Time-spent Predictor* (TSP[3]) and the *Most Popular Location* (MPL[4]), predict a user's future behaviour using only the current temporal context.

3.5.1 Most Popular Location (MPL)

The Most Popular Location (MPL) finds the location which the user visited most often. Given a time interval t, the (*MPL*) method ranks each user's visited locations based on their historical popularity. By restricting the prediction to t, as in the TSP model, MPL learns from the user temporal behaviour. For example, a user who frequently visits a shopping mall at a particular time interval of the day is likely to visit the same shopping mall during the same interval as opposed to visiting the most popular location visited in their entire mobility history.

In order to predict the location that a user u will be visiting next during the time interval t, we compute the location l, which has the highest probability of visit, amongst

[3]Time-spent Predictor
[4]Most Popular Location

the locations that were previously visited by the user u during the interval t. If $L = \{l_1, l_2, l_3, \ldots, l_n\}$ denotes the set of locations that u previously visited during the same time interval t, and $V = \{v_{l_i,j} \mid v_{l_i,j}$ is the jth visit made to location $l_i \in L\}$ denotes the set of all visits made to the locations given in L, then the probability of visit can be described by the following equation.

$$Pr(Location = l_k|t) = \frac{|V_{l_k}|}{|V|}, \tag{3.1}$$

where V_{l_k} denotes the set of visits made to the location l_k ($V_{l_k} = \{v_{l_k,1}, \ldots, v_{l_k,m}\}$), $V_{l_k} \subset V$, $l_k \in L$ and $k \in \{1, 2, 3, \ldots, n\}$.

The MPL model computes a probability of visit ranked list of all the locations visited during t and returns the *top-k* locations.

3.5.2 Time-spent Predictor (TSP)

The Time-spent Predictor (TSP) finds the location, where the user spent most of his/her time. It utilises the time-spent at each visited location as a basis for predicting the next location. For example, a user who spends more time, on average, at a shopping mall at a particular time interval of the day is likely to visit the same shopping mall during the same interval as opposed to visiting other locations.

In order to predict the location that a user u will be visiting next during the time interval t, we compute the location l, which has the highest average time-spent amongst the locations that were previously visited by the user u during the interval t. If $L = \{l_1, l_2, l_3, \ldots, l_n\}$ denotes the set of locations that u previously visited during the same time interval t, and $W = \{w_{l_i,j} \mid w_{l_i,j}$ is the duration of the jth visit made to location $l_i \in L\}$ denotes the set of durations of all visits made to the locations given in L, then the average time-spent at an observed location l can be computed by using the following equation.

$$TSP(l_k, t) = \frac{\sum_{x \in W_{l_k}} x}{\sum_{y \in W} y}, \tag{3.2}$$

where W_{l_k} denotes the set of durations of visits made to the location l_k ($W_{l_k} = \{w_{l_k,1}, \ldots, w_{l_k,m}\}$), $W_{l_k} \subset W$, $l_k \in L$ and $k \in \{1, 2, 3, \ldots, n\}$.

Property	Open challenge data set
Number of users	38
Number of user-days	8154
Average number of locations per user	89

Table 3.1: Properties of the Nokia MDC Open Challenge data set [71].

The TSP methods compute a time-spent ranked list of all the locations visited during t and returns the *top-k* locations.

3.6 Evaluation

3.6.1 Data Set

For the evaluation of the proposed approach we utilised the data set which Nokia released for its mobile data competition in 2012 [71]. To create the data set, Nokia launched the Lausanne Data Collection Campaign [71] which had nearly 200 participants and lasted for about two years. The collected data was divided into several parts where each challenge in the competition was allocated a separate part of data (see § 1.6.1.1). The data part we used (originally used for the Open Challenge) consisted of data collected from the mobile phones of 38 users. It had the actual raw location data including GPS coordinate recordings and WLAN[5] for all the users. It was rich and ideal for testing the models proposed in this chapter particularly the comparison between the collective and the one-model-per-user approaches. It is important to note herein that the proposed models in this chapter can be *directly* employed with any similar data set. The Nokia data set used herein was the only data set available to us at the time of conducting the experiments presented hereafter.

A summary of the properties of the Nokia data set can be seen in Table 3.1.

3.6.2 Experiments Design

To evaluate the proposed models, for each user, we organise the data into a sequence of days based on the time in which the user visited each of the different locations. We then carry out the following steps:

1. For each user, we divide the daily trails into two sets: a large set containing the first α % of the total number of trails. This set is used for the model training and the

[5]Wireless Area Network

remaining *(100-α)%* of the data set is used for testing. To evaluate our proposed models, the *alpha* value used to divide the data was 80 which is a common value chosen by other researchers [38]. If a user u had a hundred trails of movement in the evaluation data set then 80 trails will be utilised for training the prediction model and the remaining 20 trails will be used for testing the trained model.

2. For the one-per-user-model, we create a suffix tree for each user utilising the data given in the training set. For the collective model, we use the data from the union of the different users' training sets to create a single suffix tree. The data identifying the different users is modified so that there is only one single anonymous user associated with all the locations data encoded in the tree. Assuming that the evaluation data set contains the three users u_1, u_2 and u_3, and α % is 80%, a one-per-user suffix tree model for the user u would be built using only the 80 trails given in his/her training data set. However, to construct a collective suffix tree model we need to employ the trails given in the training data sets of two or more users such as the users u_1, u_2 and u_3. In such a model all the trails utilised in the training are regarded as being generated by a single anonymous user, and thus we can neither distinguish between the actual users nor identify the trails belonging to any of them.

3. For each daily trail data from the test set, we compute search-trails, of a maximum length n, using the following sliding window technique: Let T be the daily trail to be tested and assume a window of size n, we extract the n first locations from the trail T, match against the suffix tree and try to predict the $(n+1)^{th}$ location in T. We then slide the window to include the locations from the 2^{nd} to $(n+1)^{th}$ and attempt to predict the $(n+2)^{th}$ location and so on. The locations contained in the sliding window make, what we call, a search-trail. A search-trail of size n is identified as a rank n search-trail. To test our approach, the maximum search-trail's length used was 5, which is a reasonable choice for the number of historical visits required for querying the proposed models.

   Example: Suppose that the tree, shown in Figure 3.1, represents the mobility history of a user u and the trail "BB" gives the sequence of the most recent visits that were made. To predict the location that u will visit next, we present the trail "BB" to the tree which produces the candidate locations 'A' and 'C' - the location 'C' has a higher probability as opposed to location 'A', and thus most likely to be the next location of visit.

3.6.2.1 Error Measurement

To evaluate the accuracy of the proposed models, we use two well known metrics: the *Mean Absolute Error* (MAE) and the *Root Mean Square Error* (RMSE) [13], which are standard methods for measuring the average inaccuracy associated with a set of model-produced predictions. They essentially measure the difference between the predictions made by the observed model and the actual values being predicted. In the context of next location prediction, we have slightly different interpretation to MAE and RMSE as opposed to how they are normally interpreted in systems such as those which predict a user's likes or dislikes for items or products. MAE can be interpreted as the mean rank of the correct predictions while RMSE provides a measure of spread of the ranks of the correct predictions. MAE and RMSE can be utilised to determine the size of the list of predicted locations that includes, on average, the correct next location. For example, if the MAE value 1.5 and the RMSE value is 0.5 , one can suggest the top two predicted locations which are most likely to include the correct next location. By suggesting the top k locations on the basis of the MAE, we are actually saying that, on average, the top-k are likely to include the correct next location if it was one of the locations which the user had followed in the past.

In the proposed approach, the possible next locations are ranked 1 to r according to their probability of visit. Assuming that the highest ranked location was the one followed, then we compute the absolute error score for an individual prediction as *(r-1)*. For example, if the rank of the correct next location in the ordered list of candidate locations was 5, then the error will have the value 4. For n predictions the MAE, as shown in Equation 3.3, is the average of the n individual scores:

$$MAE = n^{-1} \sum_{i} (r_i - 1). \tag{3.3}$$

As a special case, if the location that was followed had probability zero in the suffix tree, i.e. it does not appear as a next location, then we take it to be in the last position, r. In such a situation, we assume that the list of suggested locations has the length equal to the maximum branching factor of the suffix tree (i.e. the maximum number of leaves per branch). Our choice to use the maximum branching factor as opposed to the average was purely motivated by the fact that we wanted to increase our confidence that the list of suggested locations will include the correct next location.

To compute the RMSE, the errors are squared before they are averaged. Consequently, if the squared error score for an individual prediction is *(r-1)*^{2} then for n predictions, the RMSE, as shown in Equation 3.4, is the square root of the average of the n squared error scores. Therefore, RMSE gives a relatively high weight to large errors, and is thus most useful when large errors are particularly undesirable. This implies that, in general, extending the list of suggested locations on the basis of the RMSE value, is likely to give a better chance for the correct next location to be included in the list, as opposed to when MAE is used.

$$RMSE = \left[n^{-1} \sum_{i} (r_i - 1)^2 \right]^{-\frac{1}{2}}. \tag{3.4}$$

3.6.2.2 Hit and Miss Score (HM)

To compute the Hit and Miss Score (HM), we count a correct prediction (a *hit*) as 1 and an incorrect prediction (a *miss*) as 0. We then divide the number of hits by the total number of predictions made, as shown in Equation 3.5. In the current context, HM can be interpreted as the probability of guessing that the next location visited by the user was the one with the maximum probability.

$$HM = n^{-1}h. \tag{3.5}$$

where h is the number of hits and n is the total number of predictions made.

3.6.3 Results

To develop our proposed approach, we benefited from an exploratory investigation about the users' activity during the different days of the week. We discovered a significant variance in the average number and the places visited by the users during the weekend period compared to that of the weekdays period, as shown in Figure 3.2. This was the key motivation for building separate models for each class of data.

To benchmark the performances of the proposed models, we compare the results of the proposed suffix tree models against the performance of MPL. The idea is that for a good prediction performance, the values of the three metrics: MAE, RMSE and the HM score, produced by the proposed suffix tree models, should be better than the ones produced solely by MPL. In an ideal scenario, the values of the MAE and RMSE produced by the suffix tree models should be significantly smaller than the ones produced by MPL whilst their HM scores should be, by a large amount, greater than the HM scores obtained by

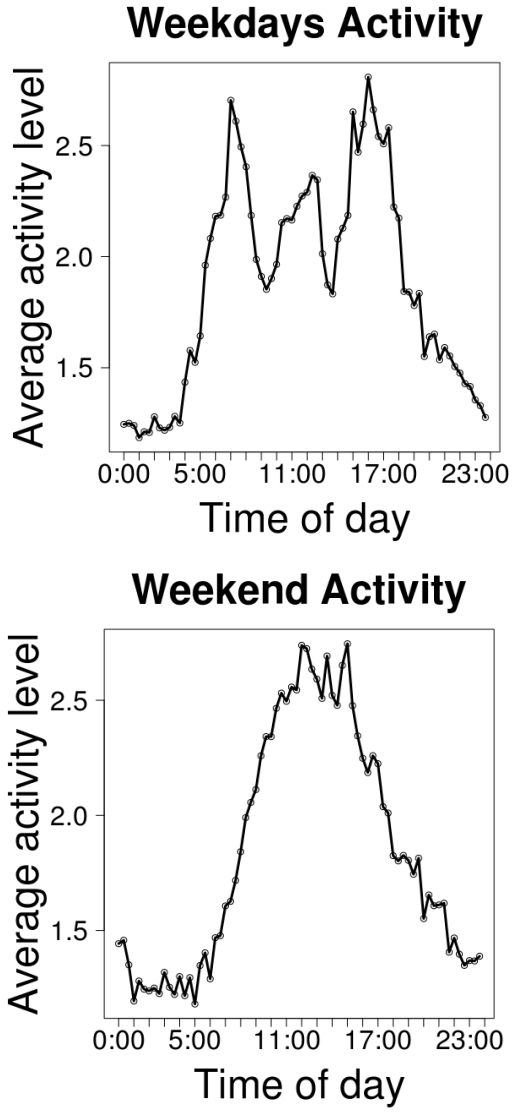


Figure 3.2: The overall daily, weekdays and weekend average activity which is computed as the ratio of total number of visits made to the number of visitors.

testing MPL. Should the results produced by the proposed models be significantly worse from this ideal scenario, the system should simply make the predictions on the basis of the user's most popular visited locations, and avoid using the more computationally expensive collective or the one-per-user suffix tree models.

3.6.3.1 Target Locations with Visiting History:

Tables 3.2 and 3.3 show the prediction results for target locations that the observed user had seen in the past; for the collective model, any previous visit to the target location could have been made by any of the users and not necessarily the observed user. The best overall HM score was produced by the collective suffix tree ($CST^6{}_{seq}$) model in which the sequential order of the visits was taken into consideration and the time of visit was ignored. Removing the time of visit increases the number of overlaps between the locations visited by the users and, as a result, there is a greater chance for the most frequently visited locations to be correctly predicted. The same model, when tested on the weekend data, reported similar HM scores as its sister model, CST which takes into account the time of visit. The rival one-per-user suffix tree (ST^7) model also had similar HM scores, when tested on the weekend data. However, the two models had very different MAE and RMSE values, with the one-per-user model reporting significantly superior results as opposed to the ones reported for the CST_{seq} model. What is exciting about the results of this experiment is the fact that those results achieved by the collective suffix tree (CST) model, were comparable to those achieved by the one-per-user suffix tree (ST) model. The two models had very similar HM scores; however, the collective model CST had a slightly higher RMSE and MAE results as opposed to the one-per-user ST model, which had the best overall MAE and RMSE results. The key contributing cause to this difference in the MAE and RMSE results was the fact that the CST model utilises data from multiple users, and hence, has a higher branching factor compared to that of the ST model.

3.6.3.2 Target Locations with no Visiting History:

The results for predicting target locations with no visiting history, shown in Tables 3.4 and 3.5, were the product of the MPL and the TSP temporal models which, unlike the suffix tree models, do not require history data to make predictions - due to the lack of history data, we cannot apply the suffix tree approach in this experiment (note here that the target locations with no history data only account for 25.63% of the total number

[6] Collective Suffix Tree
[7] Suffix Tree

Model		MAE	RMSE	HM
Single user model	MPL	7.8425	13.4803	0.6189
	TSP	10.0476	14.9909	0.5000
	ST	**1.1099**	**2.5131**	**0.7596**
	ST_{seq}	**7.0333**	**16.3389**	**0.7077**
Multi-user model	CMPL	231.6909	252.2456	0.1225
	CTSP	247.9568	260.5844	0.0717
	CST	**1.6975**	**3.8453**	**0.7547**
	CST_{seq}	**7.6994**	**19.2091**	**0.7733**

Table 3.2: MAE, RMSE and HM (In this experiment, all target locations given in the test set have visiting history in the training data set).

Model		Weekdays			Weekend		
		MAE	RMSE	HM	MAE	RMSE	HM
Single user model	MPL	7.5174	13.1452	0.6313	9.3373	14.9244	0.5621
	TSP	9.9434	14.8343	0.4961	10.5266	15.6904	0.5178
	ST	**1.1190**	**2.5174**	**0.7552**	**1.0639**	**2.4913**	**0.7820**
	ST_{seq}	6.8970	15.8087	0.7066	7.6598	18.5828	0.7130
Multi-user model	CMPL	234.9781	254.0477	0.1094	217.4076	244.2612	0.1793
	CTSP	247.9875	260.8331	0.0713	247.8234	259.5012	0.0734
	CST	**1.7552**	**3.9282**	**0.7493**	**1.4060**	**3.3961**	**0.7820**
	CST_{seq}	7.6942	19.0707	**0.7716**	7.7256	19.8934	**0.7820**

Table 3.3: MAE, RMSE and HM for the weekdays and the weekend periods (In this experiment, all target locations given in the test set have visiting history in the training data set).

of queries generated from the test data). It is noticeable here that all collective models scored very poorly, particularly the CMPL and the CTSP models which failed to make correct predictions except for a few random target locations.

3.6.3.3 Collective MPL and TSP Models:

To understand why the collective version of the MPL model (CMPL[8]) had very poor performance, we compared the variance between the numbers of users choosing a particular landmark as their most popular location. The idea is that for a given time period of the day, *CMPL* would have a good prediction performance, if a large number of users shared a particular landmark or set of landmarks as their most popular location(s). Consequently, across different landmarks, one would expect a high variance in the number of users sharing a landmark as their popular location of visit. Our experimentation shown

[8]Collective Most Popular Location

Model	MAE	RMSE	HM
MPL	15.7422	18.3290	0.1523
TSP	15.1946	17.9626	0.1514
CMPL	232.0867	243.1919	0.0222
CTSP	234.0202	244.1505	0.0161

Table 3.4: Model performance when no visiting history is available. In this experiment, all target locations given in the test set have no visiting history in the training data set.

	Weekdays			Weekend		
Model	MAE	RMSE	HM	MAE	RMSE	HM
MPL	15.7051	18.7425	0.1763	15.8000	17.6646	0.1150
TSP	15.0665	18.2973	0.1705	15.4067	17.3943	0.1196
CMPL	229.1405	242.6729	0.0268	236.5584	243.9774	0.0152
CTSP	232.2308	244.2362	0.0167	236.7360	244.0203	0.0152

Table 3.5: Weekdays and the weekends model performance when no visiting history is available. In this experiment, all target locations given in the test set have no visiting history in the training data set.

that the highest variance, across the different time intervals, for MPL models was 1.16; only three landmarks were shared as most popular locations and the maximum number of users sharing a landmark was seven.

A similar idea can be applied to the collective version of the TSP model (CTSP[9]), where the variance of the number of the users sharing a landmark as their most time-spent location is examined. Based on the experiments we carried out the reported highest variance, across the different time intervals, for the TSP models was 5.7431; only four landmarks were shared as most time-spent locations and the maximum number of users sharing a landmark was thirteen.

One approach to improve the performances of CMPL and CTSP, would be to cluster users according to visited locations or area of visit and then individually predict the mobility in each cluster using CMPL or CTSP. Implementation of such an approach would require identifying the cluster that the user belongs to. This may lead to the compromise of users' privacy if, for example, some of the detected clusters individually contain a single user. In such cases, the collective model effectively becomes a one-model-per-user.

[9]Collective Time-spent Predictor

3.6.3.4 Query Length:

We also studied the effect of the length of the historical movement trail used for querying the models to determine the relative loss of accuracy when such data is reduced. It is clear from the results shown in Table 3.7 that, on average, the more history the search query contained the better the prediction result is, except for the search queries with history data length equal to 2 which have slightly worse results compared to those with length equal to 1. This is true for both the *ST* and the *CST* models. It is also clear that for search queries with history data length greater than 3, which account for more than 70.34% of the total number of queries made in the test experiments, the *ST* achieved only a very small improvement over the *CST* HM score. This a strong indication that the performance of the two models are very similar.

Length of record	number of trails	%
0	555	25.63%
1	33	1.52%
2	29	1.34%
3	25	1.15%
> 3	1523	70.34%

Table 3.6: Details of historical records (search-trails) used for querying the models

3.7 Discussion

3.7.1 The Collective Model Versus the One-model-per-user

The collective model has a number of advantages over the one-model-per-user:

1. *Social prediction:* Like other data-mining methods, the one-per-user model can produce accurate predictions but only about mobility behaviours that have been previously observed. However, it has relatively poor performance in situations when novel behaviours occur. The collective model, on the other hand, may have better performance in such situations due to the fact that it is not only focused on the observed user's past behaviours but rather has a range of behaviours from multiple users, which quite often include the behaviour we are attempting to predict.

2. *Serendipity:* As a recommendation method, the one-model-per-user would always predict places that had been previously seen by the observed user. Whilst this leads to making safe recommendations, it does not help the user to discover new places

Length of historical record used for querying the model	ST			CST		
	MAE	RMSE	HM	MAE	RMSE	HM
1	1.4242	2.8551	0.6364	1.7576	3.7172	0.6970
2	1.6552	3.0850	0.6207	2.7931	4.8672	0.5862
3	1.2800	2.6533	0.6800	1.9200	4.0987	0.7200
> 3	1.0900	2.4907	0.7663	1.6717	3.8217	0.7597

Table 3.7: MAE, RMSE and HM, for ST and CST models, computed for different visiting history lengths (In this experiment, only target locations which have a visiting history in the ST training data set were used).

that they had not seen before. Due to the ability of making prediction using places seen by other users, the collective model can recommend new places that the user has not experienced in the past. (Note that this highlights the difference between *prediction* and *recommendation*.)

3. *Less sensitivity to cold start situations:* When users are newly added to the system, they normally have no mobility history that can be employed to predict their next behaviour, a condition which is known as cold-start [88]. Since the collective model makes its prediction on the basis of behaviours of multiple users, it is less sensitive to such situations compared to the one-model-per-user.

4. *Cheaper to build:* The collective model costs less to build and maintain compared to multiple one-model-per-users.

Despite their appeal, both collective and one-per-user suffix tree models share a few shortfalls which we summarise hereafter:

1. In the cases where there is no mobility history to consider for predicting the next location of visit, the one-per-user model fails to make a prediction. The collective model predicts the next location as the most likely location visited by the users incorporated in the model irrespective of whether it contains any record of the observed user ever visiting such a location in the past.

2. When the user has very low predictability (i.e. the user very often visits new places that he/she has never been to in the past), there is high probability that the model would make an incorrect prediction.

3. For some users, matching the highest ranked trail does not necessarily lead to a correct prediction.

To address these problems we propose the following respective solutions:

1. The immediate solution, not necessarily the most effective, is to use TSP or MPL when the observed user has no mobility history to consider. An alternative approach to deal with this problem would be to cluster landmarks based on their location and use the new data to build a collective suffix tree model for each cluster.

2. The collective model is based on historical mobility record of multiple users and as a result has a wider coverage in comparison to the one-model-per-user, which depends solely on a single user record of movement. Therefore, it make more sense to predict the behaviour of users with low predictability using a collective model as opposed to using a one-model-per-user, which is less likely to produce accurate predictions for mobility behaviours of such users.

3. In many cases, using shorter trails may result in the model correctly predicting the next location. Nonetheless, in our evaluation tests, we consider all candidate locations computed by using the longest search-trail as well as all the shorter trails from the same search-trail (i.e. we query the suffix tree using a search-trail of length 5, then length 4 and so on).

3.7.2 MAE and RMSE versus HM

To have a better perspective of the model accuracy, it is important to know, not only, whether or not the system is making correct predictions but also "how close" the prediction to matching the correct target location is when the system incorrectly predicts the user's next place of visit. Since the HM score is more focused on the hits as opposed to the misses, using it on its own, gives an imbalanced assessment of the prediction accuracy. Also, in many application areas, unless the HM score is very high, the predictions cannot be reliable and most probably not very useful. In the context of the next place prediction problem, achieving a high HM score is, generally, a very hard task. For example, in the Nokia MDC competition in 2012, the highest HM score achieved was 0.56 [78]. One interpretation of such a score is that, on average, the system will make approximately 4 errors, in every 10 predictions it makes. This is a highly unreliable score for many sensitive application areas, where there is little margin for making erroneous predictions. Measuring the prediction performance on the basis of the HM score alone does not comprehensively assess of the system's prediction accuracy.

A sensible alternative would be for the predictor to present a short list of landmarks that are most likely to be of interest to the user. Showing, the user, a list of landmarks to choose from would, in many cases, be preferable to acting on a single landmark prediction. In the current context of next location prediction, MAE and RMSE could be employed very effectively, to determine the length of the list of suggested landmarks which would, most likely, include the correct next location to be visited by the observed user.

The main criticism of any of the three metrics is that, using one on its own may not give sufficient assessment of the prediction accuracy. A thorough evaluation of the proposed prediction algorithms suggests that algorithms optimised for maximising the HM score do not necessarily perform similarly when measured with the MAE and RMSE. The experimental results show that improvements in HM score often do not translate into improvements in MAE and RMSE values. (See the experiment results in Tables 3.2 and 3.4.) With several factors to consider, using the three metrics: MAE, RMSE and HM together for evaluation purposes, is more likely to give a clearer picture of the prediction accuracy. Striking a balance between the values of the three metrics is a key element in getting a good evaluation of the overall prediction accuracy.

3.8 Summary

In this chapter, we investigated the predictive power of the collective model and the one-model-per-user approaches. We showed how the two approaches have very comparable prediction performances particularly when previously seen behaviours are available to make inferences from. We also showed that only a short record of mobility history is required in order to make relatively accurate predictions about the future behaviours of users. We examined the effect of the length of this record and the relative loss of accuracy when reduced data samples are used. It was clear that as the length of the historical record increases the the models prediction accuracy improves.

We presented an alternative approach that allows for collaborative prediction and has the potential to overcome the one-model-per-user's weaknesses such as the inability to deal with novel behaviours.

We evaluated our proposed approach using the error metrics MAE and RMSE, and showed that they can be utilised to determine the top-k landmarks which are most likely to

include the correct location to be visited next by the observed user. We also investigated the merits of HM, also know as *the success ratio*, in evaluating the models prediction performance. On the basis of the experimental results we argued that using HM on its own is insufficient to assess the prediction performance. We also demonstrated that using the three metrics: MAE, RMSE and HM together for evaluation provides a better view of the models' prediction accuracy.

Chapter 4

Mobility Fingerprinting

4.1 Overview

We define a mobility fingerprint as a profile constructed from the user's historical mobility traces. We propose an algorithm for building such a profile, and collect a sample of fingerprints from the publicly available Nokia Mobile Data Challenge data set [70]. We find that users have unique mobility fingerprints, i.e. they can be distinguished from one another. Furthermore, we find that an observed mobility trail can be associated with the fingerprint of the user to whom the trail belongs, i.e. a user can be identified by his/her movements. Here, we argue that in order to successfully identify individual users on the basis of their recent mobility history, it is imperative that a rich historical record about the movement of those users is maintained. Although it is possible to construct a minimal fingerprint while preserving its uniqueness, in the interest of user Identifiability, the richer the fingerprint the more accurate it is in identifying the correct user from a short record of observed movements. We also propose a method for constructing location fingerprints and we demonstrate how accurate such profiles can be in predicting users' future places of visit.

4.2 Introduction

Since the advent of the Internet, we have developed many forms of online identities. However, because of the continuing advancement in pervasive computing and data connectivity, perceptions about an individual's identity are changing rapidly. In user mobility, mobile devices can often be uniquely identified by the MAC address, or the "phone number".

Although such identification methods are very successful in distinguishing between individual devices, they usually pose considerable privacy challenges. For example, even when anonymisation techniques are applied and private information is removed form a data set of mobility traces, it may still be possible to associate the cleaned data with the correct individual by using outside information. The study presented in [111], which demonstrates how realistic such a compromise of privacy can be, the top locations of visit were used to re-identify the data of mobile phone users. Setting aside the sensitive privacy issues, such identification methods also cannot be employed for inferring and analysing the dynamics of human mobility. Therefore, we have been investigating the possibility of constructing a dynamic method of identification using mobility data which, for each individual user, as shown later in this chapter, possess measurable variations that make it suitable for *'mobility fingerprinting'* [32].

As shown in [111] and [98], mobility data, which usually contains detailed space and time information, can be exploited to predict accurate personal information about the movements of those people whose mobile devices generated the data. Knowledge of such information does not only give valuable insight into human mobility behaviour, but can also be of interest to a wide range of systems that benefit from accurate identification of individuals and their future locations of visit [89]. For example, this work is particularly concerned with the Identification of mobile users within the context of location prediction and recommendation. Indeed in such a context, finding a distinct set of data that makes the individual *unique* is not the key point. It is much more useful to have a rich profile that, in addition to being unique also reflects the individual's interest in terms of the places that they visit and the activities that s/he undertakes. Such a profile clearly offers a distinct advantage where it allows grouping together individuals with similar interests and tastes. The ability to create such groupings is the foundation upon which collaborative prediction and recommendation systems are developed. Furthermore, in such a context, using a mobility profile that is built from a small set of unique locations is most likely to have poor accuracy when employed to identify the user from observed movements. Restricting the profile to include only a set of unique locations would mean that any observed movements, based on shared locations with other users, would have poor similarity to such a profile; hence a successful identification of the correct user is less probable.

The rest of the chapter is organised as follows: In the next section, we present the research questions addressed, the contributions made to the field and the methodology used

for constructing the proposed fingerprinting method in this chapter. In Section 4.4, we introduce our proposed method for constructing the mobility fingerprints. In Section 4.6 we propose the idea of location fingerprinting and how it can be applied to the next location prediction problem. In Section 7.7, we describe the experimental setup and the testing methodology for the proposed methods. We evaluate our approach by collecting fingerprints from mobility traces of 38 individual users from the publicly available Nokia Mobile Data Challenge data set [71], which we describe in Subsection 4.7.1. Finally, in Section 5.7, we discuss the *serendipity* feature of the proposed location fingerprint model as well as the problem posed by compressing fingerprints of adventurous users - those that are not mainstream users.

4.3 Problem Definition

The underpinning motive behind the proposed mobility fingerprinting method, discussed herein, is to provide a platform for the development of more effective context-aware mobile prediction systems. Within the scope of this chapter:

1. We investigate whether the trails generated from users' mobility traces have sufficient measurable variations which allow for fingerprinting of movements of those users to whom these traces belong, i.e. can we create a unique profile from the user's record of historical movement.

2. Assuming that the users have different mobility fingerprints, this chapter examines the *identifiability* of the correct user from an observed mobility trail, i.e. having built a unique profile for each user, we then examine whether we can associate a short trail of observed movements with the unique profile of the correct user who generated the short trail of movements.

3. We investigate the effect of the length of the user record of the temporal locality used to correctly identify the user, and the relative loss of accuracy when reduced data samples are provided so as to establish the exact trade-off involved.

4. Focusing on the individual user, the chapter examines whether the size of the fingerprint can be reduced while retaining *identifiability*, and to this end it attempts to find a minimal fingerprint that can be employed to correctly identify the user from a short record of observed movements.

5. It also investigates whether the proposed fingerprinting method can be extended to create unique profiles for landmarks (the terms 'location and 'landmark are used interchangeably in this chapter), user activities or even temporal units such as days of the week. It examines whether such fingerprints can be used for location prediction, and to this end we demonstrate how the location fingerprints can be successfully employed in predicting the location that an observed user will be visiting in the future. Herein we refer to the definition of the next place prediction problem, which has been described in a range of other research works such as [31, 38].

4.3.1 Contributions

This chapter makes the following contributions, backed up by empirical evidence:

1. Although the term *fingerprint* has been around at least since 2013 [101], we present *the mobility fingerprint*, which is a profile constructed from a user's historical mobility traces, for predicting the user's future mobility behaviour. We propose an algorithm (see § 5.3.2) for building such a profile.

2. We demonstrate that users have unique mobility fingerprints, i.e. they can be distinguished from one another. Furthermore, we demonstrate that an observed mobility trail can be associated with the fingerprint of the user to whom the trail belongs, i.e. a user can be identified by his/her movements. Herein, we demonstrate that in order to successfully identify individual users on the basis of their recent mobility history, it is imperative that a rich historical record about the movement of those users is maintained. We also show that the richer the fingerprint is the more accurate the identification of the user from observed movements.

3. We demonstrate that the proposed fingerprinting method can be used to create unique profiles for landmarks and by successfully applying it to the Next Location Prediction problem, we demonstrate that such profiles can be a very useful tool for location prediction.

4.3.2 Methodology

We followed a specific algorithm to construct the *the mobility fingerprint*, the details of which will be discussed in Section 4.4. In brief, this algorithm takes the following steps:

Step 1. We start by detecting the different locations (or landmarks) visited by the users.

Step 2. Using the detected locations and their time of visit, we compute the mobility trails of each of the users. Each trail is represented as a sequence of n-grams (The definitions of *trail* and *n-gram* are given later in Subsection 4.4.2).

Step 3. We use the computed trails to create the fingerprints which we describe in detail in the next section.

4.4 Identification of Mobile Users Through Their Mobility Fingerprints

4.4.1 Detection of Visited Locations

We utilise the sequence of geographic coordinate recordings given in the raw data to discover the locations visited by the users. We consider here an area of concentration of GPS points as a single location (or landmark) of visit. The idea is that over time and as the user travels about in his daily routines, some of the visiting patterns to these locations would have been captured in the aggregate of these coordinate recordings. To detect these patterns we apply the following algorithm:

Step 1. We apply the K-means algorithm [110] to the GPS data points in order to create disjoint clusters.

Step 2. Compute the centroid of each discovered cluster and verify that its member points are at-most *'r'* meters away from its centroid.

Step 3. If one or more points are at distance greater than *'r'* from the centroid of the cluster, we apply (1) and (2) to the points in the current cluster for further clustering. This process continues recursively for each sub-cluster until every member point is at-most *'r'* meters away from the centroid of the cluster it is associated with.

Step 4. Each discovered cluster forms one of the locations visited by the users (Locations that have a number of visits below a certain threshold are considered noise and are thus ignored).

Note that the distance between two GPS points is computed by using the *Haversine* formula [90]. Also for our experimentation purposes, we restricted *'r'* value to 100 meters, i.e. all the visited locations are 100 meter radius. The times of visit to any of the discovered locations are the same times of visit of the individual member points belonging to the detected location/cluster.

An alternative to the proposed clustering algorithm applied herein would be to use divisive hierarchical clustering [106] which, in practice, may not work if the number of data points to be clustered is very large - it is computationally prohibitive to explore all clustering scenarios when the number of data points is very large; consequently, many proposed hierarchical clustering methods employ heuristics in order to divide the points into different clusters which can lead to unreliable results [106]. Also many proposed hierarchical clustering methods seem to be narrowly focusing on dividing the points into clusters without considering alternative partitioning possibilities - once a decision has been made to divide a large cluster into smaller ones, there is normally no mechanism for changing such a decision [106].

4.4.2 Computing n-grams to represent trails

The proposed model relies on the use of *mobility trail*, which we define as the sequence of recordings, of the temporal and spatial information, of all the visits that a user makes in a day. Let $T = t_1 , t_2 , \ldots , t_n ,$ be a mobility trail where $t_i , 1 \leq i \leq n$, represents the locations visited by the observed user. Let us also define an *n-gram* as a contiguous sequence of n visits contained in T. Given the trail T we can produce a set of all the *n-grams* contained in it as shown in the following example, in which $n = 3$.

Example: Given a trail $T = (a, b, a, b, a, b, a)$, the *Tri-grams* (i.e. *3-grams*) associated with it are given in the set $\vartheta_{T,3} = \{(a, b, a), (b, a, b)\}$. Alternatively, we can create a set of pairs where each pair record an *n-gram* together with its number of occurrence. In our example, this produces the set $\beta_{T,3} = \{<(a, b, a),3>, <(b, a, b),2>\}$ which can be represented as the multiset (or bag) $\{(a, b, a), (a, b, a), (a, b, a), (b, a, b), (b, a, b)\}$.

4.4.3 Mobility Fingerprinting

4.4.3.1 Definitions

Definition 1. *Mobility Fingerprint:*

A *Mobility Fingerprint* is a stochastic model developed from the mobility traces of an observed user to capture his/her specific movement patterns and to allow for his/her identification. The details of constructing such a model are thoroughly described in Subsection 4.4.3.2.

Definition 2. *Mobility Fingerprinting:*

Mobility Fingerprinting is the process of constructing a mobility fingerprint for an observed user.

4.4.3.2 Computation of Fingerprint

In order to distinguish users based on the information given in their mobility trails and also to be able to determine whether or not a particular trail belongs to a specific user, we propose an algorithm that relies on *fingerprinting* the user's movements. In theory, the mobility fingerprints of any two users should be different and the corresponding users of any two fingerprints must certainly be different as well. In practice, we shall see that there is only a very small probability that two different users have the same mobility fingerprint.

In order to compute the mobility fingerprint of an observed user, we perform the following procedure.

1. Select a suitable value for n (the size of the *n-grams*).
2. For each of the user's mobility trails T, we compute the set $\beta_{T,n}$ which contains all the possible *n-grams* (see the example given in Subsection 4.4.2).
3. We compute the super set $B_{u,n} = \bigcup_{i=1}^{k} \beta_{T_i,n}$ which contains all the *n-grams* from all the mobility trails (T_1 , T_2 , ..., T_k) of the observed user u.

When constructing the *fingerprint*, in addition to computing the *n-grams*, it is very useful to compute the $(n-1, n-2, \ldots, 1)$-*grams* for each trail. For example, if $n = 3$ then for each user's trail in the database we compute the *Tri-grams*, *bi-grams* and *uni-grams*. Consequently, we define a user's mobility fingerprint f as the set of all *grams* given in $\bigcup_{i=1}^{n} B_{u,i}$. Note that the value of n cannot be larger than the size of the longest mobility trail, and is in most cases much smaller.

4.4.3.3 Constructing a Unique Fingerprint

In sensitive domains such as crime-scene forensic investigations, employing techniques such as biological fingerprints and DNA sequences have become the standard methods for identification despite the fact that mistakes resulting from such methods can have substantial consequences. In human mobility, the uniqueness of the user's mobility fingerprint is related directly to the original purpose for which the mobility fingerprint was proposed, i.e.

to identify users and their mobility behaviours within the context of mobile prediction systems. Contrary to other types of fingerprints such as our biological fingerprint and DNA sequences, a user's mobility fingerprint does not have to be unique to fulfil this purpose. Nonetheless, to determine whether users have distinct fingerprints, we compare the similarity between the fingerprint of an observed user and the fingerprints of all the other users in the database. The importance of computing the similarity here lies in the fundamental role it plays in determining the *separability* of fingerprints. There may also be an application in clustering users together based on their fingerprints, e.g. for collaborative filtering like recommendation [44].

4.4.3.4 Fingerprint Uniqueness

Let F be the set of all fingerprints, in a database db, and let s denote a function for computing the similarity between two fingerprints where $s(f_a, f_b)$ of the two fingerprints, f_a and f_b fall between 0 and 1. If the value of $s(f_a, f_b)$ is close to 1, then f_a and f_b are said to be *inseparable* (i.e roughly the same). Consequently, we define the two fingerprints f_a and f_b to be *separable* if $s(f_a, f_b) << 1$.

Uniqueness

The uniqueness of a fingerprint is tightly related to its *separability*. If λ, which is a small value between 0 and 1, denotes a separability threshold, then f_a is said to be *unique* if $\forall$ $f_i \in F$, $s(f_a, f_i) \leq \lambda$, i.e. λ is the minimum similarity between any two fingerprints to be considered identical.

4.4.3.5 Similarity Computation

Jensen-Shannon Divergence

Let p = $\{p_i\}$ and q = $\{q_i\}$ denote the n-grams' probability distributions obtained from the fingerprints of the users a and b, respectively. The divergence, d, between the two users' fingerprints can be obtained using the *Jensen-Shannon Divergence* (JSD[1]) which is a nonparametric measure of the similarity between two distributions [34].

The intuition when comparing the fingerprints with one another is that we are comparing like-for-like, i.e. the two fingerprints are, good representations for the two distributions from which the n-grams originated, hence a suitable choice for computing the uniqueness

[1] Jensen-Shannon Divergence

of fingerprints is the *JSD*, which is a symmetric version of the Kullback-Leibler Divergence (KLD[2]) based on Shannon's entropy [46].

Jaccard Similarity

As shared *n*-grams between two fingerprints are more of interest than the contrary, the Jaccard coefficient, which is also known as the Jaccard measure [21, 91, 96] is our baseline method for computing the similarity between two such sets, i.e. the fingerprint and the observed trail which we view here as two sets of n-grams. We define the *Jaccard* measure by

$$Jaccard(f_a, f_b) = \frac{|f_a \cap f_b|}{|f_a \cup f_b|}, \tag{4.1}$$

where f_a and f_b are the fingerprints of user a and user b, respectively.

Example: When computing the *Jaccard* similarity between the multisets $x = \{a,\ b,\ b,\ c\}$ and $y = \{a,\ a,\ b,\ b,\ b\}$, the intersection counts a only once and counts b twice, so its size is 3. The size of the union will be the total of the sizes of the two multisets which is equal to 9. Hence, the *Jaccard* similarity between x and y is 1/3.

Weighted Jaccard Similarity

In addition to the *Jaccard* method, we also use the *Weighted Jaccard (WJaccard)* [95, 104], which we compute as follows:

$$WJaccard(f_a, f_b) = \frac{\sum_{w_i \in f_a, f_b} min(\gamma(w_i))}{\sum_{w_i \in f_a, f_b} max(\gamma(w_i))}, \tag{4.2}$$

where $\gamma(w_i)$ denotes the number of occurrences of the *n*-gram w_i.

Example: To compute the *Weighted Jaccard* similarity between the multisets x and y (see the previous example), we first compute the minimum number of occurrences of *a, b and c* (*1, 2 and 0*) and their maximum number of occurrences (*2, 3 and 1*). Then the *Weighted Jaccard* similarity is computed as $\frac{1+2+0}{2+3+1}$ which is equal to *0.5*, in this case.

4.5 Identifiability

A key question, which automatically arises when designing a mobility fingerprint, is how effective such a method in identifying the correct user from an observed trail of movements.

[2]Kullback-Leibler Divergence

To address this question, we compute the distance between the observed trail and the fingerprints of all the users in the database. The fingerprint which produces the smallest distance value is the one that, most likely, belongs to the user who generated the observed mobility trail. However, in contrast to the like-for-like comparison performed when verifying the uniqueness of fingerprints (i.e. computing the distance between fingerprints), an observed trail is only a sparse representation of the distribution of the mobility trails from which it was drawn, and consequently cannot be used for a like-for-like comparison with a fingerprint which have enough information to form a good distribution about the mobility behaviour of the observed user. In such a situation it would be more appropriate to use the *Kullback-Leibler Divergence (KLD)*, defined in equation 4.3, to measure the distance between the observed trail and the fingerprint [11]. Note here that we also employ the non-probabilistic symmetric measures, the *Jaccard* and the *Weighted Jaccard*, since they both compute the similarity between two multisets regardless of whether, or not, the multisets are good representatives of the distributions from which the *n*-grams were drawn, namely

$$KLD(\tau, f) = \sum_{x \in X} p_{\tau}(x) log_2 \frac{p_{\tau}(x)}{p_f(x)}, \tag{4.3}$$

where τ and f denote the observed trail and the user's fingerprint respectively. $p_{\tau}(x)$ and $p_f(x)$ denote the observed trail and the fingerprint probabilities for an *n*-gram $x \in X$ and X is the set of *n*-grams obtained from the test-trail.

4.5.1 Finding the Correct User

In order to speed the search for the correct user, we employ a simple heuristic that relies on the popularity of the locations amongst the different users who visited them. Before we compute the similarity between the observed trail and a user's fingerprint, for each *n-gram* found in the observed trail, we estimate the popularity as the count of users with occurrences of the same *n-gram* in their fingerprints. For example, if an *n-gram* occurs in *'x'* different users' fingerprints then the popularity of such an *n-gram* is equal to *'x'*. The importance of computing the popularity here lies in the fundamental role it plays in reducing the number of fingerprints for which we compute the similarity to the observed trail. The underpinning assumption here is that an observed trail is more likely to have been created by a user, whose fingerprint has occurrences of the less popular *n-grams* of the observed trail, as opposed to being created by any other user. Therefore, by focusing on those users with occurrences of those less popular *n-grams* of the observed trail in their fingerprints, we are more likely to predict the correct user to whom the observed trail belongs.

Another key advantage is that, because it is likely that we would have a small set of users, whose fingerprints have occurrences of the less popular *n-grams* of the observed trail, we are likely to have a faster search in comparison to a full search through the entire users' database. In fact, based on our experimentation on the Nokia MDC data set (see § 4.7.1), our heuristic search method can be up to 80% times faster in comparison to a full search. It is important to emphasise here that such a search procedure can be particularly useful when there is a large database of users to search.

4.5.2 Fingerprint Compression

The key idea here is to examine whether the fingerprint maintains the same degree of identifiability after compression. The importance of this investigation stems from the desire to have fingerprints that are computationally less expensive, i.e. occupy less memory space, while remaining unique amongst other fingerprints. Instinctively, compressing the fingerprint by eliminating redundant information, is perhaps the ideal method to resolving the problem (i.e. by reducing the amount of information in the fingerprint we reduce the amount of computational resources required). However, removing too much information from the fingerprint may result in poor identifiability. To resolve this conflict between uniqueness and identifiability, we propose two compression methods which attempt to reduce the size of the fingerprint while *preserving* its degree of identifiability, at the same time.

4.5.2.1 Temporal Compression

In this mode of compression the fingerprint is reduced in size by considering only the temporal aspect of the data. Each user's fingerprint is built incrementally from the training data as follows:

1. We start by using the entire training data from which we construct the user's fingerprint.
2. We decrease the data, used to construct the fingerprint, by omitting the oldest trail from it.
3. We test to see if we could identify, from the test set, the correct trails belonging to the observed user.
4. If we manage to correctly identify all the test trails, we go back to step 2.

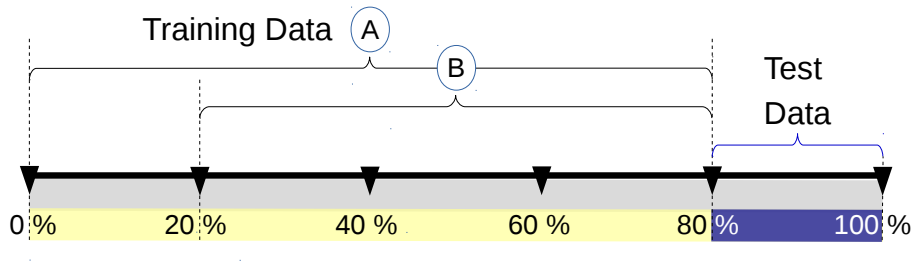


Figure 4.1: Training and testing data division (A and B show the parts of the training data used to construct the fingerprints before and after the size was reduced from 80% to 60%).

5. If we fail to correctly identify all the trails, then we conclude that the previous fingerprint had enough information to identify all possible trails belonging to the user.

4.5.2.2 Spatial Compression

In contrast to the temporal compression, the spatial compression reduces the size of the fingerprint by considering only the spacial aspect of the data. In this mode of compression, a fingerprint is built incrementally from the training data by using an algorithm similar to the one used for the temporal compression except that in point (2), we decrease the data used to construct the fingerprint by removing the locations with the least number of visits.

4.6 Location Fingerprint

Mobility fingerprinting can be very effective when accurate identification of users, locations, behaviours or activities is required. It naturally lends itself to tasks such as location prediction. In location prediction, which is used here as an example, we model the mobility behaviour involving a particular landmark as opposed to modelling the mobility behaviour involving a specific user. Similar to how we build a user's mobility fingerprint, a location fingerprint is constructed using the past mobility trails in which the observed landmark was involved. The following three steps demonstrate how such a fingerprint is built:

Step 1. Compute all the trails in which the observed landmark was visited.

Step 2. Let $T_{loc} = \{t_i\}$ denote the set of all trails of the observed landmark *loc*. For each trail in T_{loc}, compute the n-grams in the direction of the observed landmark. For example, if c denotes an observed landmark and *'abcde'* denotes a trail of

movements involving the location c, then the set of *bi*-grams computed in the direction of c will be $\{ab, ed, bc, ec\}$. For our experimentation purposes we ignore those *bi*-grams involving the observed landmark c, i.e. bc, ec.

Step 3. Compute the super set $B_{loc,n} = \bigcup_{i=1}^{k} \beta_{t_i,n}$ which contains all the *n-grams* from all the mobility trails given in T_{loc}.

4.6.1 Next Location Prediction

In order to demonstrate the accuracy of the location fingerprint, we consider the *next location prediction problem*. Given the most recent sequence of visits of an observed user, locations fingerprinting can be a powerful method for predicting where the user will go next, i.e. to determine in advance which landmark the user will visit after leaving his/her current location. A simplified version of a next location prediction algorithm that is based on mobility fingerprinting can be described as follows:

Step 1. We observe a user for a period of time to have a trail (i.e. the recent sequence of visits made by the observed user including his current location).

Step 2. Select an area from which a landmark will be predicted as the location that the observed user will visit next. For example, select an area with a radius r where the current location of the user is at the centre. *We assume here that there is an ontology for describing the type of places that the user is interested in visiting and that the user has already selected the type of place to visit, at this stage.*

Step 3. Compute the fingerprints of all suitable landmarks in the selected area. A landmark's fingerprint is computed from all users' trails involving the observed landmark - only trails that contain the observed landmark as one of the visited locations are considered in the construction of the fingerprint.

Step 4. Compute the similarity between each landmark's fingerprint and the current user's trail.

Step 5. Using the computed similarity scores, create a ranked list of landmarks from which we predict the top-k locations as potential places that the user is most likely to visit next.

The assumption here is that the similarity between the fingerprint of the correct landmark and the trails involving the same landmark will be higher compared to the similarity between the same trails and the fingerprints of other landmarks.

4.7 Evaluation

4.7.1 Data Set

We evaluate our proposed fingerprinting algorithm using a portion of the data set which Nokia released for the *open challenge* of the mobile data competition in 2012 [71]. The characteristics of the entire data set of the competition, including the details of the different portions allocated for the various challenge tracks have been previously described in § 1.6.1.1. A description of the portion of data used in this chapter is previously described in Chapter 3, namely in Subsection 3.6.1. It is worth noting here that the Nokia MDC data set is publicly available and well known for its good quality of the data. We acknowledge that it is relatively small in comparison to other data sets but sufficient for proof of concept.

4.7.2 Experiments

4.7.2.1 Testing for Uniqueness

To test for fingerprints' uniqueness we perform the following procedure.

1. We choose an experimental similarity threshold value λ (We wish λ to be as small as possible). For the uniqueness of the users in the Nokia data set, we find 8.6% an experimentally best choice, i.e. by choosing 8.6% as the minimum similarity between any two fingerprints to be considered identical, every fingerprint in the data set was found to be unique.

2. We compute the similarity between each fingerprint and the other fingerprints in the database.

3. If the maximum computed similarity value, over all the fingerprints in the database, is less than λ, we can conclude that the fingerprints are unique.

4. If the maximum computed similarity value is greater than or equal to λ, we conclude that fingerprints are not unique at the chosen threshold. In such case, we may need to consider a different value for the threshold. The appropriate choice would be to use the maximum computed similarity value as the new threshold.

4.7.2.2 Identifiability Tests

To identify the correct user from an *observed* mobility trail, it is natural that we model this task into a multi-class classification problem [86]. We divide the data into training

and test sets where either part A or part B, depending on the percentage used for training, is used for training and the remaining data is used for testing (see Figure 4.1 for training and testing data partitions). We use the training set to create the users' fingerprints while the test set is used to produce the trails which we will attempt to identify the correct users that originally created them. If M denotes the list of fingerprints, i.e. classes ($f_k \in M$ where $k \in \{1, \ldots, m\}$), created from the training data partition and N denotes the list of test trails ($t_i \in N$ where $i \in \{1, \ldots, n\}$) where each trail belongs to one of different classes f_k. Given a test trail, we compute the similarity between the test trail and each of the fingerprints in M and produce a ranked list of similarity values. To identify the correct user to whom the trail of movements belongs, we simply select the fingerprint that produces the largest similarity value.

Precision and Recall

Let tp denote the *true positives*, which represents the cases when the algorithm make the correct classification, i.e. the trail is associated with the correct fingerprint. Also let fn denote the *false negatives*, which represents the cases in which the algorithm does not associate a trail with its correct fingerprint, i.e. the trail is rejected while it should have been associated with the fingerprint. Also let fp denote the *false positives*, which represents the cases where the algorithm accepts a trail to be associated with the fingerprint, to which the trail does not belong. Considering the aforementioned notation, we can use the procedure hereafter to compute Precision and Recall:

1. We compute the similarity between each fingerprint in M and the test-trail t.
2. We create a ranked list of users on the basis of the computed similarity values. If r denotes a ranking threshold, then a positive result occurs when the similarity value of the correct fingerprint, to which the test-trail belongs, has a ranking less than or equal to r. In this case, we have a *hit* and tp will be increased by 1 and fp will be increased by $r - 1$.
3. If the similarity value of the correct fingerprint is ranked greater than r, we have a *miss* and fn will be increased by 1.

For a given threshold value r and a set of test trails N, the computation of precision

and recall proceeds as follows:

$$precision = \frac{\sum tp_{ij}}{\sum tp_{ij} + \sum fp_{ij}} \tag{4.4}$$

$$recall = \frac{\sum tp_{ij}}{\sum tp_{ij} + \sum fn_{ij}}, \tag{4.5}$$

where $\sum$ stands for $\sum_{(i \in M)(j \in N)}$.

Identifiability Coefficient (IC)

In order to fine tune the fingerprint for best *identifiability*, we have to weigh between the following competing factors: (i) the cost of incorrect classification of a trail, (ii) the cost of not classifying one at all and (iii) the ranking threshold at which the classification is made, (i.e. the maximum allowed ranking for the correct result in order to have a hit). Our approach is to find the ranking threshold value θ that maximises, pointwise, the *area under the precision and recall curve* [15]. This is the point at which we have the least combined cost of miss-classifying a trail and not classifying one at all. We refer to the ranking threshold value θ that produces the maximum area under the precision and recall curve, as the *Identifiability Coefficient (IC*[3]*)*. To compute the *Identifiability Coefficient*, for every threshold value, we compute the Euclidean distance [19] between the curve and the point (1,1). The *Identifiability Coefficient* takes the value of the threshold with the smallest Euclidean distance.

4.7.2.3 Next Location Prediction Tests

To evaluate the accuracy of the location fingerprints when applied to the next location prediction problem, we organise the data into a sequence of days based on the time in which the user visited each location. We then implement the following procedure:

1. For each user, we divide the daily trails into two sets: a training set containing the first α % of the total number of trails, and a test set containing the remaining *(100-*α*)%* (e.g. $alpha = 80\%$).

2. For each landmark, we create a fingerprint using the data given in the training set.

[3]Identification Coefficient

3. For each daily trail data from the test set, we compute search-trails, of a maximum length n, using the following sliding window technique as described in [31]:

   Let T be the daily trail from the test set, and assume a window of size n, we extract the n first locations from the trail T, match against the fingerprint and try to predict the $(n+1)$ location in T. We then slide the window to include the locations from 2 to $(n+1)$ and attempt to predict the $(n+2)$ location and so on. The locations contained in the sliding window make, what we call, a search-trail (the maximum length we used was 4).

Success Ratio

We evaluate the prediction accuracy using the *Success Ratio* which is a popular measure of prediction accuracy in domains such as Web usage mining [14] and user mobility prediction [38]. It is usually defined as the ratio of number of successes to the total number of attempts. However, when an observed system is generating a set of predictions, the *Success Ratio* is used to measure the proportion of times that the target is among the predicted set of items [14]. We find the latter definition more suitable for evaluating the results discussed in this chapter.

4.7.3 Results

4.7.3.1 Do Users Have Unique Fingerprints

For the experiments designed to test the uniqueness of the users' fingerprints, we compute a similarity matrix for all the users in the database in order to determine whether, or not, the fingerprints of different users are *separable*. One simple approach to find out would be to verify whether the maximum similarity between an observed user and the other users in the database is less than some agreed *threshold*. If that is the case, we conclude that the observed user has a unique fingerprint. Consequently, if the the maximum similarity computed across all users in the database was less than the agreed *threshold* we conclude that all users have unique fingerprints. Note here that due to the similarity measures used (*JSD*, *Jaccard* and *Weighted Jaccard*) the results may vary from one experiment to the other and consequently careful consideration is required for choosing the similarity threshold value. As shown in Table 4.1, when the similarity is computed using *Quint-grams*, the maximum similarity reported between any pair of different fingerprints is less than 8.6%, i.e. the fingerprints are completely *separable* (unique) at a similarity threshold as small as 8.6%.

Similarity Method	Statistic	The size of gram used in the similarity computation				
		Uni	Bi	Tri	Quad	Quint
Jaccard	Avg	0.0669	0.0168	0.0062	0.0027	0.0017
	Min	0	0	0	0	0
	Max	0.8768	0.4080	0.2519	0.1480	0.0856
	StDev	0.0979	0.0336	0.0157	0.0084	0.0057
WJaccard	Avg	0.0635	0.0167	0.0061	0.0027	0.0013
	Min	0	0	0	0	0
	Max	0.8904	0.8900	0.2216	0.1375	0.0825
	StDev	0.1021	0.0384	0.0150	0.0082	0.0049
JSD	Avg	0.0681	0.0171	0.0059	0.0026	0.0013
	Min	0	0	0	0	0
	Max	0.8973	0.8943	0.1842	0.1146	0.0711
	StDev	0.1225	0.0476	0.0134	0.0073	0.0043

Table 4.1: A summary of the similarity between different users' fingerprints. In this experiment, the fingerprints have been constructed from 80% of the mobility traces using the weekday data.

Similarity Method	Statistic	The size of gram used in the similarity computation				
		Uni	Bi	Tri	Quad	Quint
Jaccard	Avg	0.0964	0.0249	0.0094	0.0041	0.0019
	Min	0	0	0	0	0
	Max	0.8715	0.3673	0.1971	0.1149	0.0738
	StDev	0.1259	0.0405	0.0183	0.0095	0.0055
WJaccard	Avg	0.0784	0.0237	0.0093	0.0040	0.0019
	Min	0	0	0	0	0
	Max	0.4725	0.3232	0.1846	0.1098	0.0719
	StDev	0.0926	0.0375	0.0177	0.0093	0.0055
JSD	Avg	0.0728	0.0223	0.0091	0.0041	0.0020
	Min	0	0	0	0	0
	Max	0.4240	0.2579	0.1678	0.1081	0.0713
	StDev	0.0835	0.0335	0.0161	0.0087	0.0051

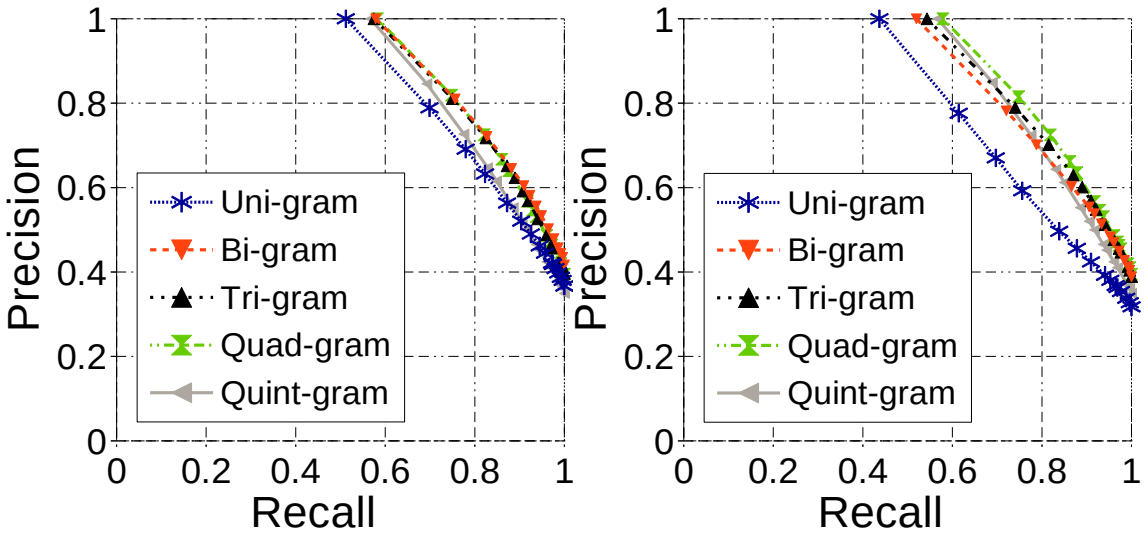
Table 4.2: A summary of the similarity between different users' fingerprints. In this experiment, the fingerprints have been constructed from 60% of the data using temporal compression.

Similarity Method	Classification Measure	The size of gram used in the similarity computation				
		Uni	Bi	Tri	Quad	Quint
Jaccard	Precision	0.7888	0.8101	0.8106	0.8189	0.8461
	Recall	0.6988	0.7559	0.75	0.7480	0.6929
	IC	2	2	2	2	2
WJaccard	Precision	0.7761	0.7820	0.7899	0.8154	0.8495
	Recall	0.6141	0.7204	0.7401	0.7480	0.6889
	IC	2	2	2	2	2
KLD	Precision	0.8067	0.8405	0.8363	0.8341	0.7012
	Recall	0.6574	0.7677	0.7244	0.6732	0.6653
	IC	2	2	2	2	3

Table 4.3: Identification of users from their movements: In this experiment, trails of five spatio-temporal points were used to identify the users. The results reported in this table are based on the same experiment shown in Figure 4.2 where only the data from the weekdays was used, and the split for the training and testing was 80% and 20% respectively.

4.7.3.2 Uniqueness of Compressed Fingerprints

When compressing the size of a fingerprint it is imperative to verify that the new compressed fingerprint preserves its uniqueness. To ensure uniqueness between the different fingerprints after compression, we must ensure that the similarity remains below the agreed threshold value λ. Table 4.2 shows the average, minimum and maximum similarity values for temporally compressed fingerprints where only 60% of the data was used (i.e. only the training data part denoted by B was used in the construction of those fingerprints - see Figure 4.1 for data division). It is evident from the results shown that the higher the gram size the clearer the separability between the different users' fingerprints is. It is also clear that the maximum similarity (i.e. 7.4%) after the compression is less than the maximum similarity (i.e. 8.6%) between the original fingerprints, i.e. the fingerprints made from 80% of the data. The reason for this difference is that the compression of the fingerprint decreases the number of shared locations (a shared location is a location which has been visited by two or more users) between the fingerprints of different users, more than it does for the unique locations (a unique location is a location that is visited by *one and only one* user). On average, the decrease in the number of shared locations was 64.41% more than the decrease in the number of unique locations, across all users.

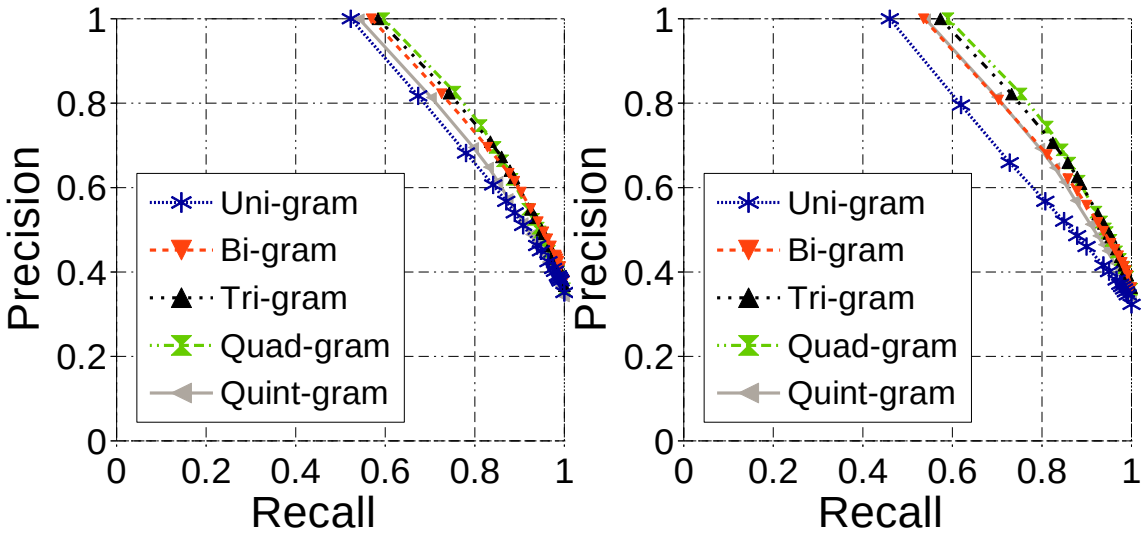


(a) Jaccard

(b) Weighted Jaccard

(c) KLD

Figure 4.2: Identification of users from their movements: In this experiment, trails of five spatio-temporal points were used to identify the users. The results reported in this figure are based on the same experiment shown in Table 4.3



(a) Jaccard

(b) Weighted Jaccard

(c) KLD

Figure 4.3: Identification of users from their movements: In this experiment, trails of five spatio-temporal points were used to identify the users. The graphs shown in this figure are based on the experiment's results reported in Table 4.4.

Similarity Method	Classification Measure	The size of gram used in the similarity computation				
		Uni	Bi	Tri	Quad	Quint
Jaccard	Precision	0.8172	0.8222	0.8236	0.8260	0.8135
	Recall	0.6736	0.7259	0.7426	0.7552	0.7029
	IC	2	2	2	2	2
WJaccard	Precision	0.7956	0.8076	0.8215	0.8219	0.8146
	Recall	0.6192	0.7029	0.7322	0.7531	0.6987
	IC	2	2	2	2	2
KLD	Precision	0.8078	0.8502	0.8776	0.8594	0.8313
	Recall	0.6861	0.7364	0.6903	0.6652	0.5983
	IC	2	2	2	2	2

Table 4.4: Identification of users from their movements: In this experiment, trails of five spatio-temporal points were used to identify the users. The results reported in this table are based on the same experiment shown in Figure 4.3 where only the data from the weekdays was used and the fingerprint was temporally compressed where only 60% of the data was used to build it. The test data was the same 20% of the data used to produce the results shown in Table 4.3.

4.7.3.3 Identifying Users From Short Trails of Movements

The Figures 4.2 and 4.3 show the precision and recall graphs which summarise the results of the experiments we carried out to identify a user from an observed mobility trail, i.e. identify a user from his/her movements. In the said figures and the tables 4.3 and 4.4 (in which IC denotes the *Identifiability Coefficient*), we find that the uncompressed *Bi-grams* KLD model, which is the best performing model in this case, has 76.77% recall and 84.05% precision. The small corresponding Identification Coefficient (IC) value 2, means that for 76.77% the observed test trails, one of the two top selected users is the correct user to whom the observed trail belongs. When the fingerprints are compressed to include only 60% of data, the reported recall and precision are 73.64% and 85.02%, respectively, where the corresponding IC value is 2. It is evident from the results of these experiments that even when the fingerprint is compressed, we can still successfully identify the correct user from a short trail of movements.

4.7.3.4 Minimum Fingerprint

A key aspect of compressing a fingerprint, ought to be about its minimum size where it can still be uniquely identified from a correct user's trail of movements. To quantify the size of the fingerprint after the compression, one has to address the following two questions:

Q 1. What is the probability of constructing a minimal unique fingerprint and what level of identifiability such a fingerprint can offer?

To address the first question let's assume that each observed user has at least one single spatio-temporal point that is unique to them. In such a situation, we can certainly build a fingerprint for each individual user utilising the single point that is unique to them. Although every fingerprint will certainly be unique amongst the other fingerprints, none will be useful in identifying the observed user when he/she starts visiting different locations as opposed to the one recorded in their fingerprint. Given the evaluation data utilised in the research, building such fingerprints is certainly possible as the minimum number of unique spatio-temporal points per user is greater than one, which is shown in Table 4.5. This means that a single unique location is enough for a user to be uniquely identified within the data set. This is, of course, consistent with the finding of the research carried out in [27] where only a small number of locations is required to uniquely identify a user. However, the problem does not lie in constructing a minimum unique fingerprint but rather, the complexity lies in maintaining an acceptable level of identifiability after the fingerprint compression, i.e. (i) how to construct a fingerprint which has acceptable level of user identifiability and (ii) how far can we compress this fingerprint size, while maintaining the ability to correctly identify the user from a short record of movements. We already dealt with part (i) when we proposed our fingerprinting method, in Section 5.3.2. The latter part (ii) will be addressed next in this section.

Q 2. How much compression a fingerprint withstands before an acceptable level of identifiability is compromised?

As we already know the threshold value for the maximum similarity between uncompressed fingerprints, we can only perform further compression if the maximum similarity across the newly compressed fingerprints remains less than the known threshold value (λ as defined in Subsection 4.4.3.4). The other key factor to take into consideration here is the *Identifiability Coefficient* and its associated levels of *precision* and *recall*. It will only make sense to perform further compression if the Identifiability Coefficient and its associated levels of precision and recall remain at acceptable levels. To this end, we designed a number of experiments in which the trails of each user were split into training and test sets. Figure 4.3 shows the precision and recall for the different grams'

Type of Location	Average	Min	Max	Standard Dev.
Unique	292	2	1544	346
Shared	587	1	4198	950

Table 4.5: The distribution of unique and shared locations. A shared location is one which was visited by two or more users and a unique location was visited by only *one* user.

sizes when the fingerprint was temporally compressed to include only 60% of the data. The 20% of the data used for testing the uncompressed fingerprint in the previous experiment remained unchanged in this experiment. By comparing the results shown in Tables 4.3 and 4.4 we find that when only the top 2 results are considered (i.e. the Identifiability Coefficient is 2), the winning Bi-gram KLD model recorded a 3.13% reduction in recall. In fact, the average of the reduction in recall across all the Bi-gram models was 2.63% and the overall reduction across all models was 1.08%. Although further research may be required here to find out the true implications on identifiability when the fingerprint is compressed, it is nonetheless evident, from the results shown in Tables 4.3 and 4.4, that the further the fingerprint is compressed the poorer the identifiability will be.

4.7.3.5 Uniqueness versus Identifiability

The experiments conducted to identify the users from their movements, before and after the fingerprints compression, show that the richer the fingerprint the more accurate in identifying the correct user. Based on the results given in Table 4.3 and Table 4.4 when the top 2 ranked locations are considered, there is an average decrease in recall of 0.63% across the different fingerprint models after compression. Also by examining the similarity measures defined in equations (4.1), (4.2) and (4.3), it is clear that the separability of the fingerprint improves with the increase of the number of the unique locations (a unique location is a location visited by one and only one user) that the user visits and the number of visits they make to those locations. It also improves with the decrease of the number of shared locations (a shared location is a location visited by at least two users) that the user visits and the number of visits they make to those locations. As the majority of the users visits a fair number of shared locations, as shown in Table 4.5, there is a desire to shrink the size of the fingerprint in order to improve its separability. This raises a question about the extent to which the fingerprint can be compressed before an adverse

Rank of correct result	Fingerprint Model			
	Uni-gram	Bi-gram	Tri-gram	Quad-gram
$= 1$	0.2857	0.5079	0.5135	0.6250
≤ 2	0.4603	0.6349	0.6216	0.7500
≤ 3	0.5555	0.7460	0.7027	0.7812
≤ 4	0.6031	0.7619	0.7027	0.7812
≤ 5	0.6349	0.7619	0.7567	0.7812

Table 4.6: Predicting the next location using the historical record of the most recent visits. In this table, which shows the prediction Success Ratio, the maximum length of the user's trail used to make the prediction was four locations. The data split for the training and testing was 80% and 20% respectively. The similarity computation was based on KLD.

effect on its identifiability becomes significant. In our view, when compression is being considered, striking a balance between the fingerprint's uniqueness and its identifiability may be required. The decision of where the balancing point lies ultimately depends on the application in which the fingerprint is being used.

4.7.3.6 Next Place Prediction

The results given in Table 4.6 show the prediction of the next location using a historical record of at most four locations. From these results one can immediately observe that the longer the historical record the better the prediction is. To our knowledge, there is no like-for-like comparable result in the literature but, with 78% of the correct predictions ranked within the top 3 locations, the results seem promising nevertheless.

4.8 Discussion

4.8.1 Serendipity

A key feature of the proposed algorithm (see § 5.3.2) is that it is not restricted to predicting the next location of visit from the set of locations that the observed user has seen before. This is because the algorithm predicts the next location of visit from the nearby set of locations around the observed user's current location, which means it can potentially predict a location that the user has never seen previously. In a recommendation context, this means the algorithm has the capacity to suggest new locations that have never been visited in the past.

4.8.2 Adventurous Users

The two types of fingerprint compression, described in Section 4.5.2, are computationally expensive to maintain over time. With an adventurous user, new data is added every time the they visit new places that they have not seen before. Over time, the compressed fingerprint will grow to a point where it may require to be further compressed to reduce its size. This cycle will properly be repeated several times until the fingerprint is no longer needed. Repeated compressions for a small number of fingerprints may not pose a problem but for a large database it may introduce a real computational challenge, particularly if there is a large number of users who frequently visit new locations. One approach to decrease the number of fingerprints that require compression would be to limit the increasing total number of locations in a fingerprint, which can be controlled by introducing a decay factor for how recently a location is visited. The assumption made here is that those locations that have not been actively visited by the user, for some long time, may be removed from the fingerprint.

4.9 Summary

We proposed the mobility fingerprint which is a profile constructed from the user's historical mobility traces. We proposed an algorithm for building such a profile. We evaluated our proposed algorithm by collecting a sample of fingerprints from the publicly available Nokia Mobile Data Challenge data set (see § 4.7.1). It is worth noting here that the Nokia MDC data set is publicly available and well known for the good quality of the data in it. We acknowledge that it is relatively small in comparison to other data sets but sufficient for proof of concept.

We verified that users have unique mobility fingerprints, i.e. they can be distinguished from one another. Furthermore, we verified that an observed mobility trail can be associated with the fingerprint of the user to whom the trail belongs, i.e. a user can be identified by his/her movements. Herein we showed that in order to successfully identify individual users on the basis of their recent mobility history, it is imperative that a rich historical record about the movement of those users is maintained. We also showed that the richer the fingerprint the more accurate the identification of the user from observed movements is.

We demonstrated that the proposed fingerprinting method can be used to create unique profiles for landmarks and by successfully applying it to the *Next Location Prediction problem*. This shows that such profiles can be a very useful tool for location prediction.

Chapter 5

Presence Analytics

5.1 Overview

This chapter illustrates how aggregated Wi-Fi activity traces provide anonymous information that reveals invaluable insight into human presence within a university campus. It shows how technologies supporting pervasive services, such as Wi-Fi, which have the potential to generate vast amounts of detailed information, provide an invaluable opportunity to understand the presence and movement of people within such an environment. It demonstrates how these aggregated mobile network traces offer the opportunity for human presence analytics in several dimensions: social, spatial, temporal and semantic dimensions. These analytics have real potential to support human mobility studies such as the optimisation of space use strategies. The analytics presented herein are based on recent Wi-Fi traces collected at Birkbeck, University of London, one of the participants in the Eduroam network [42, 105].

5.2 Introduction

The increasing advancement in wireless technologies together with the widespread use of new generations of faster and more powerful mobile devices has greatly improved the ability of people to access information while moving about in their daily lives. This increasing accessibility to digital information has the potential to generate vast amounts of detailed information, providing an invaluable opportunity to study different aspects of presence and movement behaviours of people within a given work or study environment. Furthermore, the increasing level of connectivity to information sources is affecting our environments and the way they operate and therefore, it is essential that we build real-time

monitoring systems as well as theoretical frameworks to understand how people's presence and its dynamics reshape the structures of our environments. With such measurements put in place, we can discover hidden patterns of behaviour at both the collective and at the individual user levels, thus increase our understanding about people's presence, and in turn, improve our ability to make informed decisions when we plan for our environments.

In the context of this chapter, we analyse the mobility traces generated by users accessing the wireless network at Birkbeck, University of London. Birkbeck is a full participant of Eduroam[1] (Education roaming), a WLAN service developed for the international education and research community, that gives secure, world-wide roaming access to the Internet [42, 105]. The findings reported in this chapter are the result of the analysis carried out on the Eduroam access traces, for the period from the 1st of October 2013 to the 10th of April 2015. In comparison to most data sets used in previous Eduroam based studies, this data set is larger in size with respect to its number of users as well as the number of days it spans [1, 77].

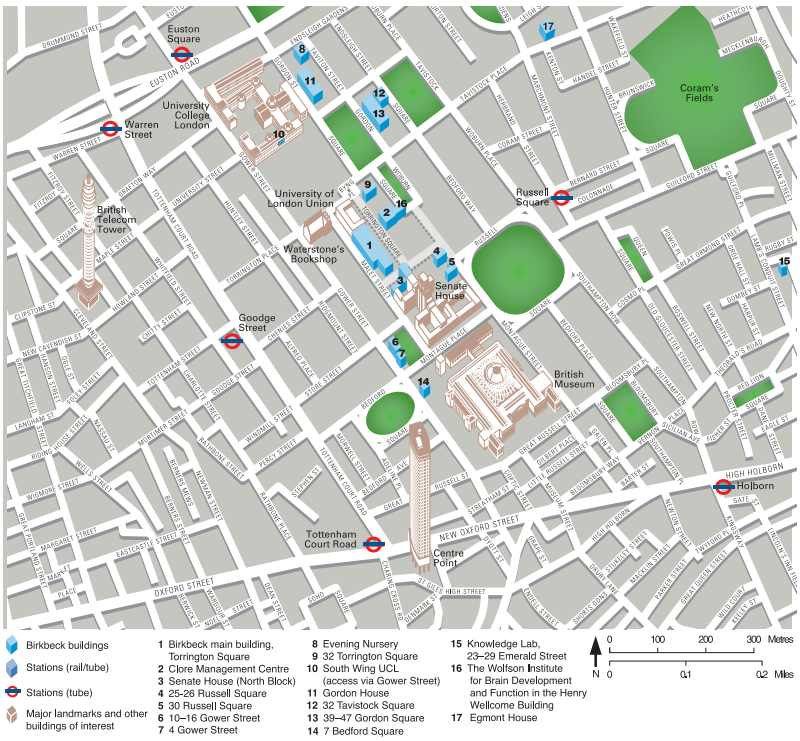


Figure 5.1: The location of Birkbeck's Bloomsbury Campus in central London.

[1]Education Roaming

The rest of the chapter is organised as follows: In Section 5.3, we present the motivation behind the research presented herein and the contributions made. In Section 5.4, we describe the essence of presence analytics, giving the definitions of the concepts used in this work as well as describing the metrics used for discovering the different patterns of human presence contained in the data. The description of the data set used, the analysis carried out and the evaluation of results are provided in Section 5.6. We give a comprehensive discussion in Section 5.7 and conclude with a short summary of the findings of the work presented herein.

5.3 Problem Definition

The prime motivation of the research presented in this chapter is to provide an insight into the human presence in a learning environment such as a university campus. Within the scope of this chapter, we examine the human presence patterns corresponding to the four data aspects, namely the social, the temporal, the spatial and the semantic.

5.3.1 Contributions

This chapter makes the following two contributions:

1. It presents a comprehensive analysis of the human presence within a university campus. It provides a thorough analysis with respect to the different types of pattern contained in the data, namely the social, the spatial and the temporal patterns, and the semantic underpinning, giving an insight into how people presence shapes the dynamic structure of such an environment. For each of these pattern types: the social, the spatial, and the temporal patterns, and the semantic underpinning, we define a list of metrics, which we utilise to interpret the observed behaviour captured in the data. Although there are numerous previous works investigating the network usage of users in WLANs [7, 57, 66], there was no attempt to analyse the four aspects of human presence in one study - a succinct summary of some of these research efforts, which generally concentrate on characterising the network usage utilising one or two of the data aspects at most, was provided in the Critical Review chapter, namely Chapter 2.

2. To our knowledge the analysis provided in this chapter is based on the most current data set - compared to data sets used in previous Eduroam research - and thus reflects the behavioural trends in Wi-Fi usage in a university setting. With the

exception of the data set used in [26], the data set used in the research herein is much larger in size compared to data sets used in previous Eduroam based studies [77]. A further distinguishing property of the evaluation data is the fact that a larger proportion of users come from other universities in the vicinity as opposed to being affiliated with Birkbeck, University of London.

5.3.2 Methodology

We investigate the four patterns of the human presence, namely the social, spatial, temporal, and the semantic underpinning, contained in the mobility data of an observed environment, giving an insight into how people presence shapes the dynamic structure of such an environment. We utilise a combination of data analysis methods to extract these patterns, where each group of methods target a specific pattern type. For example, for the extraction of the temporal patterns of revisits, we deploy time series analysis.

5.4 The Essence of Presence Analytics

5.4.1 Definitions

Definition 1. *Presence Analytics:*

Presence Analytics is defined as the collection and the analysis of mobile data in order to find meaningful patterns about people's presence within a given environment.

Definition 2. *Event:*

An *event* is defined as a group of one or more users, i.e. devices, connecting to the network from a particular location within a given time interval.

Definition 3. *Revisit:*

A *revisit* is defined as the appearance of a user at a previously visited location or site.

Definition 4. *Duration of Stay:*

Duration of Stay is the length of time that a user spends at a given location before moving to another location.

Definition 5. *Pattern of Event:*

A *pattern of event* is defined as a time series of occurrences of a given event, associated with a given time, e.g. evening or weekend pattern.

Within the context of this chapter, we rely on wireless network traces to gain information about human activity, in order to unravel the dynamic structure of the environment. Based on WLAN traces, activity patterns can be compared through time and space to reveal the dynamic structure of the observed environment. Presence analytics allows for classifying the locations within a given WLAN environment, into functional clusters based on the time-line of human activity, providing valuable insights into the actual space use patterns within that environment. It provides new ways of looking at the structure of a given environment from a real-time perspective based on dynamic up-to-date records of human presence.

5.4.2 Data Sessionisation

Unless explicitly recorded, it is usual that WLAN access data do not include session duration information, i.e. the length of time an individual was using the network service to access information. Session duration information is essential for the type of analysis presented in this chapter, where it plays a key role in computing space occupancy duration at an observed environment. It is important to highlight here that in this thesis, we prefer the term *duration of stay* over the term *session*, as it encompasses the social, the spatial and the temporal information required for presence analytics.

The data set we use in this work does not include the duration of stay information. Nonetheless, this data set has *sufficient* related information that can be utilised to compute approximately the duration of stay.

When the user authenticates more than once within the same day, we compute their *approximate duration of stay* at the observed location using the sequence of authentications made by the user on that day. Practically, we apply a threshold based method utilising the length of the interval between the times of the user's authentications. This method is described as follows:

Using the timeline of authentications, we compute an ordered list of all the authentications made by the user. We also select a lower threshold value as well as an upper value - e.g. 1 and 110 minutes - and then apply the following procedure to compute the duration of stay:

1. We assume a duration of stay equals to the lower threshold value, i.e. one minute, for those users with a single authentication record per day.

2. If the user authenticated from the same location several times, and the difference between the times of two consecutive authentications is below the upper threshold value, then we assume that the user was present for the entire interval between those two consecutive authentications and thus the duration of stay is computed as the difference between the times of the two authentications as shown the the following example.

   Example: Let the lower and the upper threshold values be 1 and 110 minutes, respectively. If a user u has two consecutive authentications at 10:00 and at 10:30. Since the difference between the two authentications of the user u is 30 minutes, which is below the upper threshold value (i.e. 110 minutes), we compute the duration of stay as the total time between the two authentications, and thus it is equal 30 minutes.

3. If the user authenticated from the same location more than once and the difference between two consecutive authentications is larger than the upper threshold value, then we assume that the user's presence at the observed location was interrupted and consequently we compute two separate *duration of stay*: one duration that includes all authentications made before the interruption, and another that includes all those that took place after it.

4. If the user moved to a different location after a single authentication, we assume that his duration of stay at the previous location is equal to the lower threshold value.

5. If the user has made a sequence of authentications from the same location, and the difference between the times of two or more consecutive authentications is below the upper threshold value, we compute the accumulated sum of duration of stay for these consecutive authentications.

The lower and upper threshold values referred to above, i.e. 1 and 110 minutes, adhere to the minimum and maximum session duration found in the literature [77].

5.5 Discovering Meaningful Patterns about the Human Presence

We distinguish between four pattern types of the human presence in this work; each group corresponds to one of the aspects of the data, namely social, spatial, temporal and semantic aspects. These four pattern types are discussed hereafter.

5.5.1 Social Patterns

The patterns discussed in this category characterise the human presence from a social perspective. To examine such patterns, we utilise a collection of metrics which measure the influence of the social behaviour in the data. Some of these metrics capture where the users come from. For example, in this work we utilise some of the metrics to investigate the distributions of the users' study or work affiliation to see what social characteristics can be extracted. The metrics address the following questions:

1. What are the institutions that the users are affiliated with.
2. Which affiliations take the top ranks in terms of the number of users.
3. How does the number of Birkbeck users compare to the numbers of users from other institutions particularly *during the evening*.

5.5.2 Spatial Patterns

The patterns discussed in this category capture the spatial view of the human presence. The metrics used herein analyse the users' behaviour from a spatial perspective. For example, we study, through these metrics, how the space is used - how the different locations are used by the users. We look at the impact of the division of the space into multiple sites. We examine the patterns of revisits to individual sites as well as the individual locations within the sites. Furthermore, we investigate a number of questions relating to the use of space at Birkbeck. The metrics address the following questions:

1. Which locations tend to be used the most by the top ranked affiliations
2. What times these locations are being used.
3. What locations do Birkbeck users use the most.

5.5.3 Temporal Patterns

In this category we view the data as time series and examine it for the existence of trends and seasonal patterns. We also examine the time that the user spends at a given location.

5.5.4 Semantic Underpinning

We define the *semantic underpinning* as the reason for the occurrence of a human presence at given spatial, temporal and social context. It is an integral part of the analytic framework, presented in this chapter, for understanding the influence of the presence of humans within a given environment. This framework is based on four main questions involving the different pattern types of the human presence. We describe the social patterns as the *who* patterns, the spatial patterns as the *where* patterns, the temporal patterns as the *when* patterns, and the semantic underpinning can be described as the *why* aspect of the human presence.

In order to decipher the *why* aspect, we investigate the use of external information in giving meaning to the user's presence - in other words, we are trying to answer the question of *why* the user's behaviour was seen within the given social, spatial and temporal context. For example, why a student is seen in a specific lecture room at a given time or why a group of students and staff were seen at the coffee-shop between 12:00 and 1:00pm.

Further discussion about the semantic underpinning is provided later in Sections 5.6.6 and 5.6.11.

5.6 Evaluation

5.6.1 Data Set

5.6.1.1 Birkbeck, University of London

Birkbeck is one of the member colleges of the University of London and a major provider of evening higher education. Based on the most recent available statistics, there are approximately 12,054 students attending Birkbeck. Most of Birkbeck students are part-time, with approximately 62% of them enrolled on part-time programmes [99].

Birkbeck's Bloomsbury campus in central London is located very close to campuses of

other colleges of the University of London, such as UCL[2] and SOAS[3]. This proximity to these other campuses was naturally translated in a large amount of collaboration between these universities. As a result, Birkbeck's Bloomsbury campus is shared by thousands of academics, researchers and students from these universities on a daily basis.

Birkbeck is also one of the participants of Eduroam, a WLAN service developed for the international education and research community that gives secure, world-wide roaming access to the Internet [42, 105].

5.6.1.2 Eduroam Data

The analysis presented in this chapter is based on recent WLAN traces collected at Birkbeck (see § 1.6.1.2). The data set we used here is a snapshot of the Birkbeck's Eduroam access data for the period, from the 1st of October 2013 to the 10th of April 2015. It contains 223 locations and 167272 users, who come from 2462 institutions and departments. The 223 locations given in this data set are divided between 11 of the 17 sites of the Bloomsbury campus (see Figure 5.1).

There are four types of data in this data set: *authentication details*, *pre-proxy details*, *post-proxy details* and *reply details*. User-ID, access location, timestamp and affiliation are the basic information for each processed record. Based on these records, we designed new types of data representing the four aspects of the human presence: social, spatial, temporal and semantic aspects. The analytics presented in this work are based on this new data.

5.6.2 Experiments

We have four types of data experiments in this section. Each set of experiments is targeted at detecting the statistical distributions of one of the observed aspect of the data: social, spatial, temporal and semantic. These experiments are described as follows.

5.6.3 Detection of Social Patterns

Given that the users in the data set are socially grouped by affiliation, we wanted to find out how they are distributed across these affiliations. The experiments here are aimed

[2]University College London
[3]School of Oriental and African Studies

at detecting four different socially distributions of users: daytime, evening, weekdays and weekend.

5.6.4 Detection of Spatial Patterns

The 223 locations given in the data set are divided between 11 sites within the Bloomsbury campus. As we are interested in finding out how the space is used, many of our efforts were focused on investigating the number of people visiting these sites and the locations within them. Since we are interested in the regular patterns of space use, we considered the revisits and excluded the single visits to the sites, i.e. all visits by individuals who came to Birkbeck only once are considered an unnecessary noise.

5.6.5 Detection of Temporal Patterns

The main goal of the experiments conducted in this section is to analyse people's presence from a temporal perspective. We are interested in discovering the trends of regular presence as opposed to the occasional behaviours due to special events. Therefore, visits by individuals who only came to Birkbeck once are seen as special events, and thus excluded from this analysis in order to avoid the introduction of undesired noise. We use data decomposition to estimate the seasonal influence as well as any random noise. By removing these estimated components from the data, we can reveal the temporal trend in people's behaviour. Since the seasonal variation in the number of revisits is relatively constant over the period of time covered by the data, we consider an additive model for the decomposition, namely

$$\textit{Presence Data} = \textit{Trend} + \textit{Seasonal} + \textit{Noise}. \tag{5.1}$$

5.6.6 Detection of Semantic Underpinning

In Section 5.5, we focused our discussion on the core three dimensions of the human presence; the social, the spatial and the temporal dimensions. This was followed by a brief note about the semantic dimension and how it can potentially add meaning to the patterns linked to those three core dimensions. In this section, we study the semantic influence by using an example in which external data from the teaching timetable is utilised. In this example, we use a two-step process, which can be described as follows:

Step 1. We obtain the information about the pattern we would like to discover, from the external source.

Step 2. We analyse the data for temporal patterns that possess distinguishing characteristics and which match the information obtained from the external source.

Of course, this process can be completely reversed, where we begin with the extraction of the patterns from the data and then map or link those discovered patterns to the information obtained from the external source in order to justify the existence of these patterns.

5.6.7 Results

5.6.8 Evaluation of Social Patterns

Figure 5.2 describes how users are distributed across affiliations. It provides four different distributions: daytime, evening, weekdays and weekend. Each one of these four distributions is approximately a power law - the fitted distributions have been computed using maximum likelihood estimation (MLE) as described in [24, 47]. In these four distributions most users belong to a small number of affiliations while the many more affiliations have a relatively small number of users. Each plot shows the Complementary Cumulative Distribution Function (CCDF), which is defined as $Pr(X \geq x)$ [24], plotted on logarithmic scales.

5.6.9 Evaluation of Spatial Patterns

Each of the individual plots presented in Figure 5.3 shows the fitted distribution across all locations visited during a chosen period, e.g. Daytime. Each of them plots the distribution of number of users' revisits, with associated power law, exponential and log normal fits - plotting the data in this manner has the benefit of providing an easy comparison between a number of fitted distributions. As can be seen in the figure, i.e. Figure 5.3, the number of users' revisits are *approximately* log normal distributions across locations. This means that the further we move across the locations towards the tail of the distribution, the quicker the decrease in the number of revisits made to the location. At the thin tail of the distribution the locations are rarely visited. The fitted distributions, in Figure 5.3, have been estimated by maximum likelihood estimation (MLE) as described in [24].

5.6.10 Evaluation of Temporal Patterns

The generally constant seasonal variation is clearly visible in the bottom plots of the Figures 5.4 and 5.5 shown hereafter.

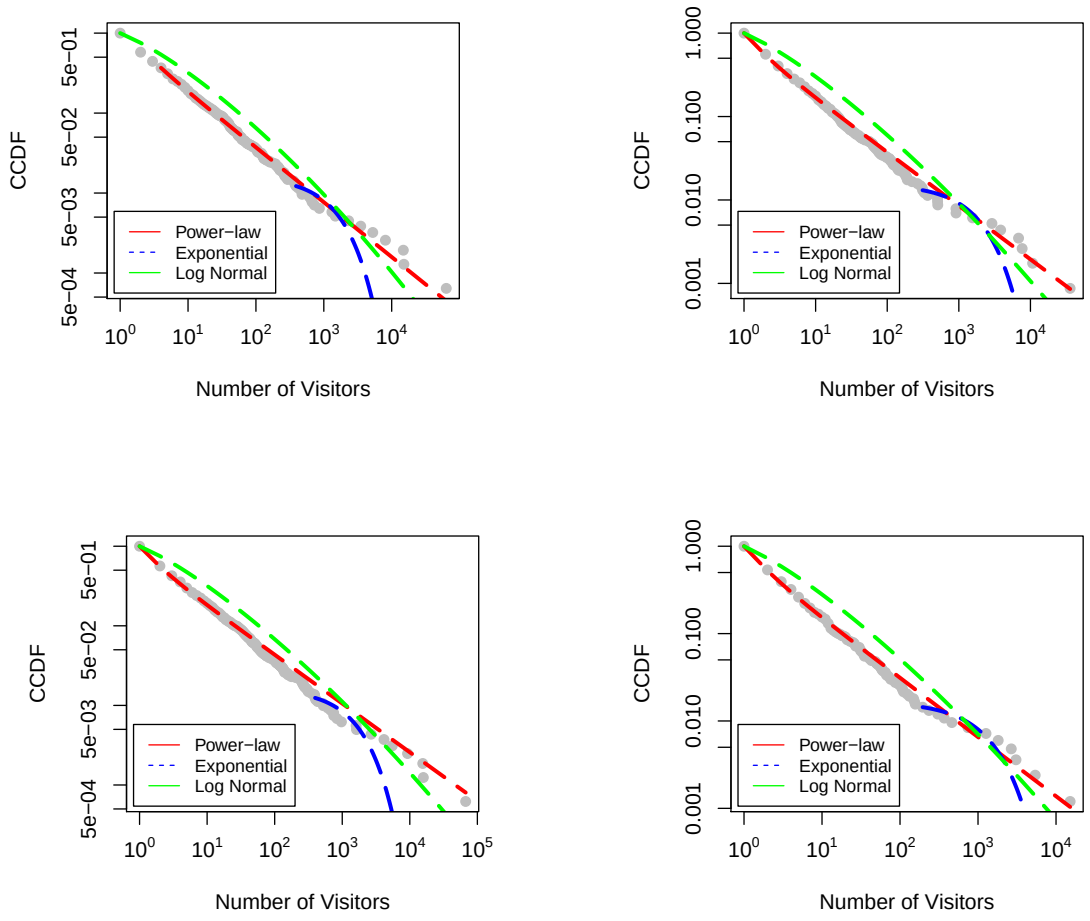


Figure 5.2: Distributions of number of revisiting users grouped by affiliation. Shown from left to right are the plots of number of revisiting users for: daytime and evening, and weekdays and weekend. Each plot shows the Complementary Cumulative Distribution Functions (CCDF) [24] and their maximum likelihood: power law (red), exponential (blue) and log normal (green) fit. Revisiting users are those who made more than one visit to Birkbeck, University of London.

5.6.10.1 Term-based Signature

We are interested in the trend, which illustrates the temporal variation in the number of revisits across the different academic terms. To extract such a trend, we divided the data into 13 weeks periods and applied an additive model to estimate the constituent behaviours such as the seasonality. Here, regular holidays such as Christmas and Easter holidays are considered seasonal events, which are captured well by the additive model (see the plots given at the lower parts of the Figures 5.4 and 5.5 which show these extracted seasonal

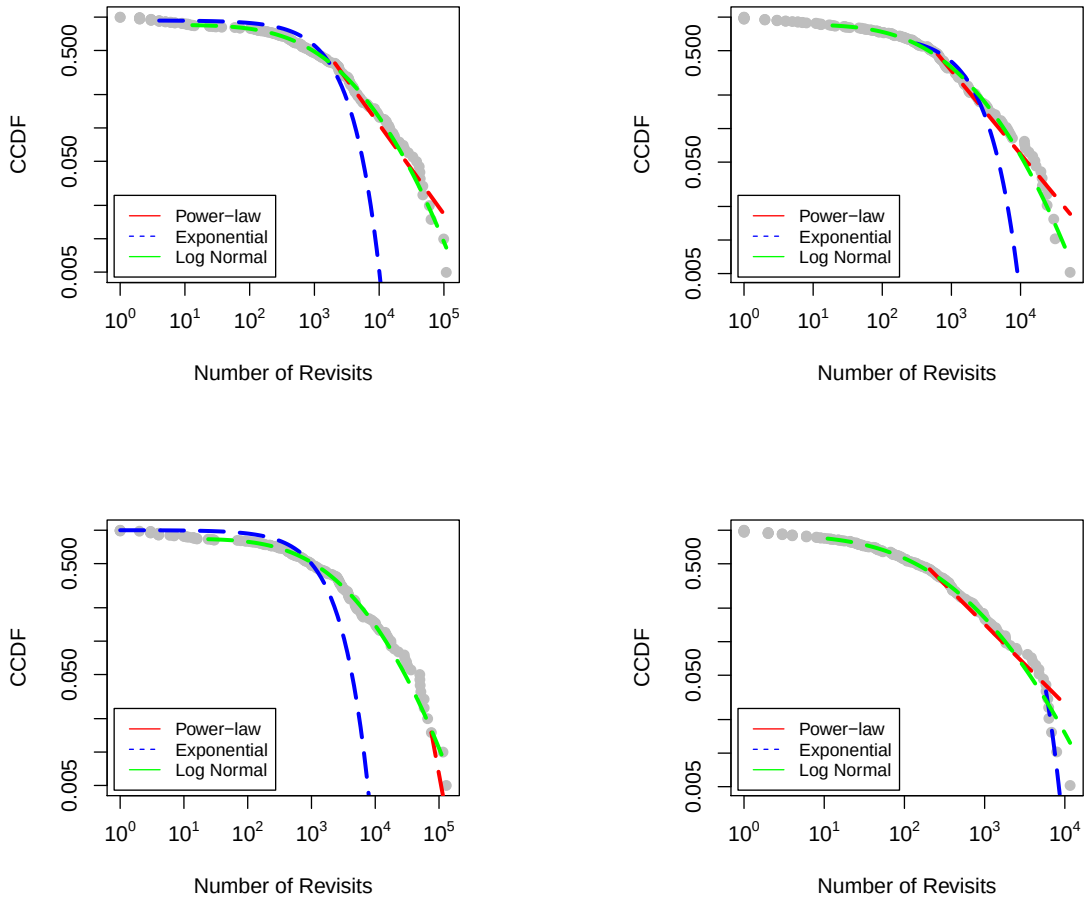


Figure 5.3: Distributions of number of revisits by location. Shown from left to right are the plots of number of revisits by location, for: daytime and evening, and weekdays and weekend. Each plot shows the Complementary Cumulative Distribution Functions (CCDF) [24] and their maximum likelihood: power law (red), exponential (blue) and log normal (green) fit. The number of revisits made to a given location is computed as the number of visits decreased by one.

events). The dipping points in the seasonality graph shown in Figures 5.4 and 5.5 can be linked to such events.

The extracted termly trends for both Malet Street and Gordon Square sites show very similar patterns. In Figure 5.5, which shows the time series analysis for Gordon Square site, we see that the estimated trend shows a decrease from about 3000 revisits in the second period to about 500 revisits in the fourth period. This decrease preceded a steady

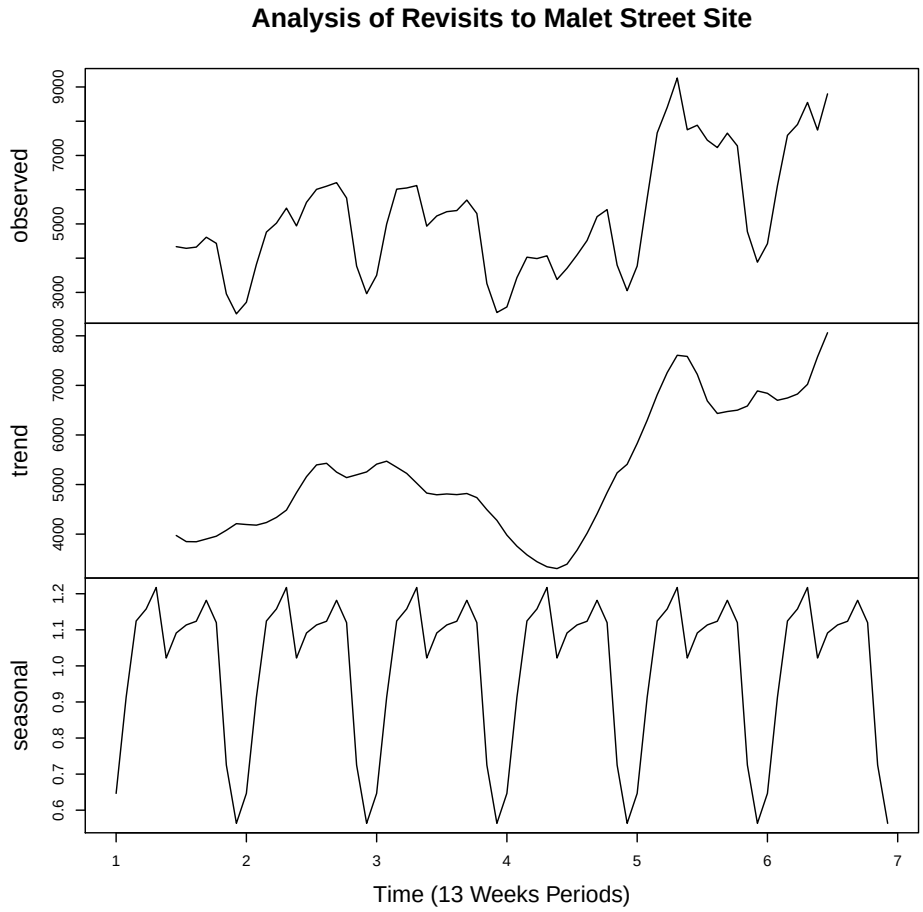


Figure 5.4: Time series analysis of number of revisits to Malet Street site. In this figure, the top plot shows the original time series in which the data is divided into 13 week periods, the plot second from top shows the estimated trend, and the bottom plot shows the estimated seasonal constituent

Figure 5.5: Time series analysis of number of revisits to Gordon Square site. In this figure, the top plot shows the original time series in which the data is divided into 13 week periods, the plot second from top shows the estimated trend, and the bottom plot shows the estimated seasonal constituent

increase to about 4000 revisits in the sixth period. In Figure 5.4, which shows the time series analysis for Malet Street site, we see almost an identical pattern to the trend shown at Gordon Square site. We see that the computed trend shows a decrease from about 5500 revisits in the second period to about 1000 revisits in the fourth period, followed by a steady climb to about 7500 revisits in the sixth period. The computation and the time series analysis carried out to produce the plots shown in the Figures 5.4 and 5.5 was performed in R [100].

In Figure 5.6, we see that the distribution of the time that a revisiting user spends, on average, at a given location is *approximately* a log normal distribution across locations. This means that the further we move across the users towards the tail of the distribution, the faster the decrease in the time spent by the user, and at the thin tail of the distribution the users spent very little time at their visited locations. The plot given in the figure, i.e. Figure 5.6, shows the Complementary Cumulative Distribution Function (CCDF), which we define as $Pr(X \geq x)$, plotted on logarithmic scales.

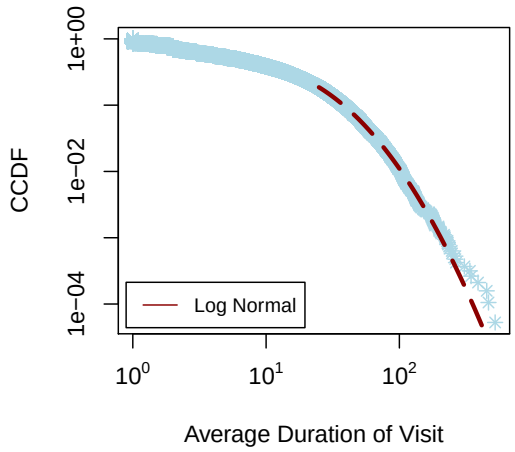


Figure 5.6: Distribution of average duration of stay constructed on a logarithmic scales. The fitted curve shown in this plot - the dash line - has been estimated by maximum likelihood estimation (MLE) as described in [24].

5.6.11 Evaluation of Semantic Underpinning

5.6.11.1 Linking Detected Learning Activities to the Teaching Timetable

By combining social, spatial and temporal information with external data, such as the teaching timetable, we can identify groups of users whose presence at a given location is primarily due to a specific semantic influence as opposed to any other reason. To demonstrate this, we conducted an experiment in which we analysed the socio-spatio-temporal patterns for one of the computer labs at Malet Street site. From the teaching timetable, we selected the XML[4] module, which had regular teaching sessions that ran from 18:00 to 21:00 every Monday in the period from the 12th of Jan to the 9th of March 2015. Socially, there was a total number of 14 individuals who attended the college for this module and had recorded traces within the data set. Note here that in order to map these patterns onto the teaching timetable we only consider the extracted patterns for the period of time that the teaching sessions cover. The result of this experiment is given in Figure 5.7, which shows the attendance of individuals from the selected group. In this experiment, the aggregate number of those who actually attended the sessions was 13. There was only one individual, who had no traces within the extracted patterns for the observed spatio-temporal context in which the sessions were taking place. Interestingly, this individual had traces in other locations when these sessions were taking place. The failure to detect the attendance of this individual can be attributed to the mobile device of this user being switched off while the session was taking place.

The example given above is a proof of concept for the influence of the semantic underpinning and the value that it adds to the other three dimensions of the human presence analytics. Of course this chapter is mainly about the core three dimensions but no doubt the semantic dimension is an important feature that provides invaluable insight into the presence of people within a given environment, and thus could not be excluded from the discussion provided herein.

5.7 Discussion

5.7.1 Analysing Human Presence Aids Planning

Although this chapter does not include how the users utilise the wireless network, in terms of the applications being used, the type of information being transferred and the

[4] eXtensible Markup Language

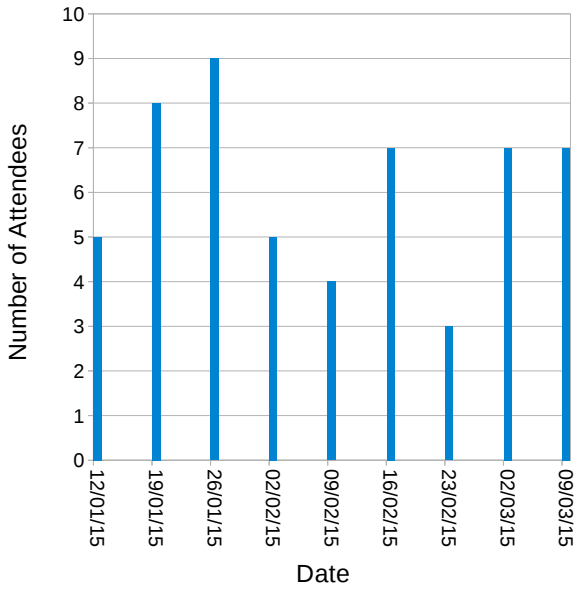


Figure 5.7: Attendance of the XML Module sessions as seen through traces of WLAN activity. The small attendance value recorded on 23/02/2015 was for the module reading week, in which the regular class session did not run.

rate in which the transfer happens, the analysis it gives already provides excellent insights about the way people interact with their environment. This suggests that reports showing accurate information about the presence and movement of people within the environment can be a very useful tool for planners when making decisions about the restructuring of the environment and how it operates. As people's connectivity to information sources becomes more ubiquitous and widespread, utilising such tool will become more common, perhaps as an additional tool to the more traditional ones such the expensive surveys and the classical static maps and drawings.

5.7.2 Data Limitation

Eduroam has the advantage that it is pervasive throughout the university and requires a single setup for authentication. Similar to most WLAN services, without registering the mobile devices with the service it is not possible to obtain any activity traces that can be linked to the users of those devices. In the experiment about the attendance of the XML sessions (discussed in § 5.6.11), a larger proportion of those who attended the class did not have traces in our data set. These individuals might have not registered with the Eduroam service and thus we could not track their activity and determine their whereabouts when the teaching sessions were taking place. This shows a limitation in utilising Eduroam for mobile devices tracking within a given environment.

5.7.3 Granular Social Groups

Social relationships is an integral part of every community and no doubt that numerous relationships and social networks exist between members of the same university community. The people at Birkbeck are no exception to this. Unfortunately, there is no explicit information, about social relationships other than the affiliation, directly available from the data set. However, through the day-to-day social activities such as lectures, seminars and regular meetings, we have strong evidence for the existence of *finer-grain* social grouping as opposed to the grouping of people provided through the user affiliation; for instance, a group of students who attend the same class is a finer-grain social group compared to the group comprising the entire community of people affiliated with University. Extracting such finer-grain grouping is a key investigation of the next chapter of this thesis.

5.8 Summary

We provide a comprehensive analysis, about the human presence within a university campus. We investigate the four types of patterns contained in the data: the social, the spatial, and the temporal patterns, and the semantic underpinning, giving an insight into how people presence shapes the dynamic structure of such an environment. Our analysis is based on WLAN activity traces collected at Birkbeck, University of London. These traces are the most recent Eduroam data in comparison to data used in other previous Eduroam research [1, 77], and thus the provided analytics reflect the current behavioural trends in WLAN usage in a university setting.

From a social perspective, our analysis reveals that the distribution of revisiting users across the various affiliations is approximately a power-law. The various patterns investigated: daytime, evening, weekdays and weekend, show that most users belong to a small number of affiliations while the many more affiliations have a relatively small number of users. However from a spatial perspective, we discovered that the users' revisits aproximately follow a log normal distribution across locations. From a temporal perspective, the extracted termly trends show very similar features of revisit. The trend generally seems to gradually decrease reaching its lowest point in August, followed by a steady climb that reaches a peak at the end of November or the beginning of December.

To demonstrate the influence of the semantic dimension we show how combining social, spatial and temporal information with external data can give meaning to a user's

behaviour. As a proof of concept, we give an example (see § 5.6.11.1) in which we utilise the teaching timetable to interpret the presence of a group of students attending a three hour regular session that took place on a weekly basis in one of the computer labs at Malet Street site.

Chapter 6

Mobile Users' Social Grouping

6.1 Overview

Utilising density-based clustering, we illustrate how granular social groups of mobile users can be detected within a university campus, using Wi-Fi activity traces. The proposed density-based clustering algorithms in this chapter, can automatically discover the learning classes, attendance data, and social groups of students who attend the same classes. For the evaluation of our proposed methods, we utilised a large Eduroam log from the case-study university (see § 1.6.1.2). We successfully detected the regular learning activities, and estimated the attendance levels over the academic term period.

6.2 Introduction

Eduroam [42, 105] and other pervasive wireless technologies, generate vast amounts of detailed information, which provides an invaluable opportunity to study different aspects of people's presence and movement behaviours within work, study or leisure environments. These pervasive technologies increase people's ability to access information, which undoubtedly affects the way the target environment operates. It is therefore essential that we build real-time monitoring systems as well as theoretical frameworks to understand how people's presence and its dynamics reshape the structures of such environments. With these measurements put in place, we can potentially discover hidden patterns of behaviour at both the collective and at the individual user levels, thus increase our understanding about human presence, and in turn, enhance our ability to make informed decisions when we plan for our environments.

The remainder of the chapter is organised as follows. In Section 6.3, we describe the goal and the contributions of the research work presented herein. Section 6.4 discusses the detection of the attendance of learning activities. In Section 6.5, we describe the social clustering of mobile users, where we investigate the socialising that occurs outside the classroom. In the Evaluation section, we give a description of the data set used for the evaluation of our proposed approaches, and present the results of the experiments we conducted. We provide a comprehensive discussion about some of the key features of the methods proposed herein. Finally, in Section 6.8, we give a summary of the chapter.

6.3 Problem Definition

The key research goal, in this chapter, is to detect granular social groups of mobile users, within a university campus, that reflect the users groupings such as those formed on the basis of attending lectures of individual modules, and those that underpin the socialising that takes place during the break-times.

6.3.1 Contributions

This chapter makes the following contributions:

1. It presents social density-based clustering methods that uses WLAN traces in order to detect granular social groups of mobile users within a university campus. The proposed clustering methods rely on the underpinning semantic context for parameterisation, i.e. utilising information from the semantic context to determine the values of the clustering algorithm parameters such as the minimum class size value, which we use to ensure that the number of individual students in any discovered social group remains within a certain range values.

2. Makes accurate estimates about the actual level of attendance of learning activities. Linking the discovered social group that regularly visits an observed location and the learning activity that takes place within the same context, will allow us to estimate the attendance level of these learning activities.

6.4 Attendance of Learning Activities

In this section, we place special emphasis on investigating the social dimension of the human presence within an academic environment, with the objective of discovering meaningful social clusters of users. In particular, we apply our proposed algorithms using the Wi-Fi activity traces that visitors leave behind as they move about from one location to another across the different sites of the chosen case-study institution - Birkbeck, University of London. Given that there is regular teaching that takes place at this university, our intuition is that we would be able to discover clusters that match the users groups formed on the basis of attending lectures of individual modules. Gaining knowledge about the social group that regularly attends a target class - the group's size and its coherence - allows us to estimate the attendance level. Furthermore, by clustering learning activities (e.g. modules) together one may be able to discover a higher level of grouping that matches the clustering formed with respect to the membership in the study programmes that the students are enrolled in.

The raw WLAN traces used in this research were collected at Birkbeck, University of London during the period from the 1st of October 2013 to the 10th of April 2015. In comparison to most data sets used in previous Eduroam based studies [1, 77], the data set containing these traces is larger in size with respect to its number of users as well as the number of days it spans.

6.4.1 Motivation

In the previous chapter, namely Chapter 5, we investigated the human presence within an academic environment and examined four types of behavioural patterns that correspond to the four different aspects of the data: social, spatial, temporal and semantic. Motivated by the findings, we set out to study more closely the social aspect of presence analytics, with the aim of gaining a better understanding of the human presence within the case-study academic site - the Bloomsbury campus of Birkbeck University of London. Based on the analysis presented in Chapter 5 there is high temporal regularity in the human presence (see the evident seasonality pattern in Figure 5.4), which can be interpreted as the visitors having preferences with respect to the visited locations. Moreover, our analysis reveals that the distribution of revisiting users across the various affiliations is approximately a power-law [24]. The various patterns investigated, i.e. daytime, evening, weekdays and weekend, show that most users belong to a small number of affiliations as can be implied

from the analysis of the distribution, shown in Figure 6.1. It is not surprising that these affiliations, which include Birkbeck College, are the ones that hold the most regular teaching and research activities across Birkbeck's sites. Furthermore, as shown in Figure 6.2, we discovered that the number of users' revisits across all locations follows a log normal distribution. The combination of these findings gives very strong indications of an underlying semantic users/visitors grouping on the basis of the learning activities that take place at Birkbeck's Bloomsbury campus in central London.

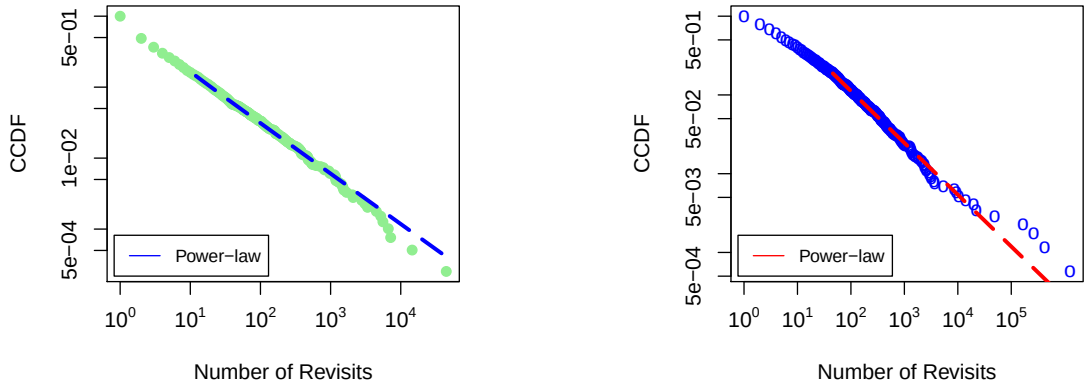


Figure 6.1: Distribution of number of users and the number of revisits by affiliation, using logarithmic scales. In this figure, the left plot shows the distribution of number of revisiting users grouped by affiliation, and the right plot shows the distribution of number of revisits made by those users. Each plot also shows the best fit line computed by maximum likelihood estimation (MLE) as described in [24].

6.4.1.1 The Intuition of the Proposed Approach

Social relationships are an integral part of every community and there is no doubt that numerous social communities exist between the people of the same university. The people at Birkbeck are no exception to this. Based on the day-to-day social activities such as lectures, seminars and other regular meetings, we have strong evidence about the existence of finer-grained relationships as opposed to the high-level social grouping inferred by the user's academic affiliation (the academic affiliation was obtained from the domain name of the user's email address). In this chapter we propose density-based clustering approaches to discover the social groups formed on the basis of these learning activities. Our choice

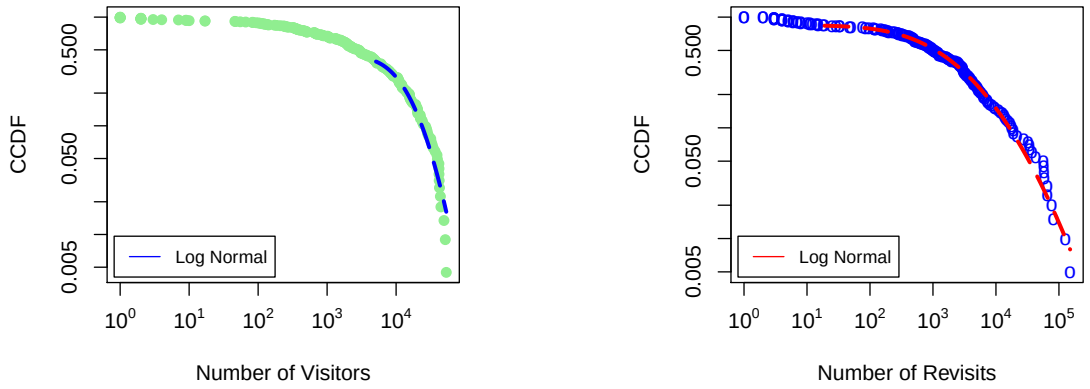


Figure 6.2: Distribution of number of users and number of revisits by location, using logarithmic scales. In this figure, the left plot shows the distribution of number of revisiting users of each location, and the right plot shows the distribution of number of revisits made by those users. Each plot also shows the best fit line computed by maximum likelihood estimation (MLE) as described in [24].

of a density-based clustering over other types of clustering methods is motivated by the semantic underpinning of the visits made to the various locations in the College. In most cases, when a location is visited, the visit is normally motivated by the desire to attend the learning activity taking place at the target location. For instance, when a student makes a visit to one of the lecture-rooms he or she is most probably doing this because they are attending a class taking place at that location.

It is important to note here that with exception to the minimum class size and the minimum attendance threshold, which we discuss in Subsection 6.4.3.1, we do not make any specific assumptions about the level of attendance of any given regular learning activity. Moreover, we do not make assumptions about the density or the variance of attendance or the shapes for the clusters that we would like to discover. The reason is that these measurements about the attendance, i.e. the density and the shapes of the social clusters of users, are partly the kind of information that we set out to discover in this research, and consequently we take into consideration an unbiased prior view about them.

6.4.2 Clustering of Users

The patterns discussed in this section are concerned with the social perspective of the human presence, in particular, the human presence with the respect to learning activities that occur across the different locations at a university campus. To examine such patterns, we utilise a collection of methods to measure the influence of the social behaviour in the data. Some of these methods capture the degree of similarity between users, while others are designed to detect the social groups of these users.

6.4.2.1 Definitions

We would like to extend the set of definitions given in Chapter 5 (see Section 5.4) by adding the following new definitions.

Definition 1. *Break-time:*

A *break-time* is defined as a time interval during the day in which the students are not engaged in a regular learning session; the definition is general to include not only breaks during a class but also before and after a class occurs.

Definition 2. *Social Coherence:*

Social Coherence is the similarity between presence and movement behaviours of two or more students.

Definition 3. *Classroom Social Group:*

A *classroom social group* is defined as a group of students who attend the same class and maintain a social connection or relationship outside the classroom. A member of such group is referred to as *a classroom friend* or just *a friend*.

Definition 4. *Noise:*

Noise is defined as the presence of mobile users within a given spatio-temporal context but this presence is not linked to the attendance of the regular weekly class session that takes place within the same context. For example, the presence of an individual in a classroom at the time when a regular weekly class session takes place, while s/he is not a member of the group of students who usually attend the observed class session, can be seen as noise.

6.4.2.2 Problem Formulation

Suppose that we have the individual users' records of revisits, of a group of users U, to a target location L. Moreover, suppose that all these revisits were made within a fixed time interval of a given weekday D for k consecutive weeks. We would like to automatically discover whether this collection of revisits represent a *pattern of events* of a learning activity that was taking place at the location L over the k consecutive weeks.

In the remainder of this chapter we use the terms *learning activity*, *class* and *pattern of events* interchangeably to refer to the same concept. Similarly, we sometimes mention users, people, students and visitors all to mean the same thing.

6.4.2.3 Distance Measure

An important question that automatically arises when we want to decide whether an observed user can be associated with a particular group of users, is how to compute the distance between the observed user and the members of a group. An equally important question to address here is how much information is required to determine a realistic value for such a distance. To answer these questions we utilise information extracted from the semantic context to inform our model about the kind of distance measure to use and the amount of information needed to compute the distance between two users' records of attendance.

6.4.2.4 Jaccard Distance

We choose Jaccard Distance [19], which we argue is a natural measure, based on the application and the data. Intuitively, the Jaccard Distance, substantially captures the difference between two records of attendance. It is defined as 1 minus the Jaccard similarity [20, 21, 91, 96], which we compute as the ratio between the intersection and the union of the two compared records of attendance [20]. The formal definition of this metric, as a function $d()$ that takes two arguments, can be given as follows:

$$d(p_a, p_b) = 1 - \frac{|p_a \cap p_b|}{|p_a \cup p_b|}. \tag{6.1}$$

where p_a and p_b represent the records of revisits of user a and user b, respectively.

Student ID.	Attendance Record	Jacc Distance		
		s_1	s_2	s_3
s_1	{*1,2,3,6,7,8,9,10,11*}	0	0.11	0.73
s_2	{*1,2,3,6,7,8,9,10*}	0.11	0	0.82
s_3	{*4,5,6,7,11*}	0.73	0.82	0

Table 6.1: An example for the term-based distance computation. The numbers given in the sets representing the attendance records, correspond to the IDs of the sessions attended by the students. The sets do not feature the sessions that the student did not attend.

6.4.2.5 Distance Computation

One of the key challenges to address when computing the distance is how much information is required to determine a realistic value. Within the context of the work presented in this section, the presence and movement of people can be highly dictated by the learning activities that takes place across Birkbeck, University of London. For example, the regular presence of students and the teaching staff in lecture-rooms is highly dictated by the learning activities that occurs in these rooms. Similar to other academic institutions, these learning activities such as lectures and lab sessions are highly dictated by the timetable, which gives the location and time allocation for the different learning activities across the academic year. Here at Birkbeck, this allocation is usually different for the different academic terms, with exception to a selection of core modules that continue to run for more than one term. Nonetheless, within the term period many people are likely to be present at the same location at the same time at least once a week. This observation was confirmed by the regularity found in the temporal patterns as shown in Figure 5.4. Based on this finding, we decided to compute the distance over the 11 week periods - each 11 week period corresponds to one of the academic terms contained in the data set described in Subsection 6.6.1.

Example: Suppose that we have three students s_1, s_2 and s_3, who attended a class c that ran every Monday for 11 weeks. We can denote the attendance of each of these students as a set representing the student's individual session attendance. In Table 6.1, which shows the Jaccard distance between the attendance of the three students, s_1 and s_2 have similar attendance while the pair s_1 and s_3 and the pair s_2 and s_3 have dissimilar attendance records. In this example, two records of attendance are considered similar to one another if they have a Jaccard distance value which is lower than 0.5.

6.4.3 Discovering Regular Classes

To explain how our proposed method successfully detects the occurrence of a class, we rely on the intuition that the visitors to an observed location, where the regular sessions of a module are delivered, naturally form a social group that meets on a regular basis over the number of weeks that the class covers. The experiments we conducted, (see § 6.6.2.1), were designed to discover such groups by performing a two-stage process, which addresses the following challenges.

6.4.3.1 Noise Reduction

Since using MAC addresses and AP location to sense occupancy is inaccurate at room granularity due to inconsistent wireless connectivity of devices and the overlap of AP coverage [75, 85], it is not guaranteed that all detected individuals at a particular classroom were there, merely to attend the learning activity taking place at that location. In order to successfully detect a regular class that takes place at an observed location, we discard from our processing the data of any individual whose total number of visits to the location was less than a *minimum attendance* threshold.

Another key factor that is closely related to level of attendance is the *minimum class size*, which is the smallest percentage of the total number of students registered for the class that must be present for a learning session to hold. Note here that the *minimum attendance* and the *minimum class size* vary between the different schools and departments within the case study university. For the proposed method evaluation, we experiment with a range of values, of those two parameters, and we discovered that by restricting the *minimum attendance* to 40% we obtain the most realistic class sizes, i.e. within the capacity of the obserevd rooms, (see § 6.6.3.1).

6.4.3.2 Attendance Coherence

Even with the noisy data being removed, we still cannot guarantee that those individuals who visited a particular location were there merely to attend the learning activity that was taking place there. It is, therefore, imperative to verify that those students taking part in the potential class are *coherent* in attending its individual sessions, over the 11 week of the academic term. A *coherent* cluster here is defined as a group of individual users that have *similar* attendance. For example, if two or more students consistently attended the same sessions of a class then they are members of a *coherent* cluster/group.

In the distance computation example given earlier (See Table 6.1), assigning the students s_1 and s_2 to the same group is likely to result in a coherent cluster, whereas grouping the students s_1 and s_3 or s_2 and s_3 together is likely to create an *incoherent* group.

In order to verify the coherence of attendance, we apply our proposed clustering method to find out whether those individuals, whose attendance satisfy the minimum attendance requirement, form a single *coherent* group - in terms of attending the individual sessions across the different weeks of the academic term period.

We emphasise herein that due to the individual pair comparison of attendance records utilised in the proposed clustering method, the students do not have to attend exactly the same sessions in order to be clustered together in the same group.

6.4.3.3 Discovering Coherent Clusters

The clustering approach we are proposing herein is based on the DBSCAN algorithm, the density-based spatial clustering of applications with noise [35], which scales well with clustering large amount of data [67]. The original DBSCAN takes two parameters, namely epsilon (a distance threshold) and minPts (a minimum number of points which is used as a density threshold). Given some data points for clustering, DBSCAN relies on these two parameters to identify density connected points in the data. It uses the concepts of direct and density connectivity to group points together forming transitive hulls of density-connected points, which yields density-based clusters of arbitrary shapes. In DBSCAN, two points are said to be *directly connected* if they are at distance less than the threshold epsilon and a point is said to be a *core point* if it has more directly connected neighbouring points than the threshold minPts. Furthermore, two points are said to be *density connected* if they are connected to core points that are themselves density connected to one another [67].

In our proposed social variant of DBSCAN, which we refer to as Social-DBSCAN, we use information from the semantic context of the human presence to inform the DBSCAN algorithm about the distance and the density threshold values, which the algorithm utilises to discover the social clusters present in the data. In particular, there are two main differences between our version of DBSCAN and the original version published in [35]:

1. The distance measure we utilise is based on the Jaccard coefficient, which as discussed earlier in Sections 6.4.2.4 and 6.4.3.2, plays an important role in capturing

the degree of coherence between the records of attendance of individuals in the same social group, i.e. it compares the total of shared attended sessions to the sum of sessions attended by either of two observed users.

2. A further important difference, which is closely related to the data set being clustered and also related to how the clustering is performed, is that the points representing the individuals who attended the learning activity, are ordered in descending order based on the individual's level of attendance, i.e. ordered by the individual's total number of attended sessions. The ordering of the points in descending order captures the idea that the higher the level of attendance the more likely that the individual is part of the social group that attended the learning activity. This is a key concept of how the clustering is performed in our proposed version of the DBSCAN algorithm [35].

To illustrate how the proposed algorithm works, let us consider the simple scenario given in the example provided in Subsection 6.4.2.5, in which the session attendance information and the similarity between the different students are given in Table 6.1. The result of applying the proposed algorithm on the scenario described in the example is that the two students s_1, s_2 are clustered together in one cluster whereas s_3 is considered to be noise - the minimum size of a cluster in this case is equal to 2 and the maximum allowed Jaccard distance, between any pair of students grouped together in the same cluster, is 0.5.

Building on the DBSCAN algorithm described in [35], the proposed clustering method, which we call Social-DBSCAN, is described in Algorithm 6.1. The two parameters *CohCoff* and *MinClassSize* represent the *coherence coefficient* and the *minimum class size* threshold values, respectively. Note here that the value of the *CohCoff* is computed as a Jaccard distance, i.e. $1 - Jaccard$, as defined in Section 6.4.2.4. From a practical perspective, the values of these parameters are heavily influenced by the context in which Social-DBSCAN is being applied. In Algorithm 6.2, we illustrate how a discovered social group is expanded.

Both Social-DBSCAN and DBSCAN have similar complexity due to the computation of the neighbourhood for every point in the data set being clustered and thus the complexity for the two algorithms is $O(n^2)$ [110].

```
1: Social − DBSCAN(Dataset, CohCoff, MinSGroupSize = 2, MinClassSize)
2: # The data set is ordered in descending order of attendance level.
3: # 'MinSGroupSize' represents DBSCAN's 'MinPts' paramater.
4: set DiscoveredSGroups ← φ
5: set SGroup ← φ
6: set VisitedPts ← φ
7: set NOISE ← φ # A variable holding all unclustered points.
8: for each (p ∈ Dataset) do
9:     if p ∈ VisitedPts then
10:        continue
11:    end if
12:    VisitedPts ← VisitedPts ∪ {p}
13:    # FindSimilarPts() returns all points similar to p, i.e. points
14:    # with Jaccard distance < CohCoff. The returned set of points is
15:    # presented in descending order according to attendance level.
16:    set pSimilarPts ← FindSimilarPts(P, CohCoff)
17:    if |pSimilarPts| < MinSGroupSize then
18:        NOISE ← NOISE ∪ {p}
19:    else
20:        set SGroup ← ExpandSocialGroup(p, SGroup, pSimilarPts, CohCoff,
21:        DiscoveredSGroups)
22:        if |SGroup| ≥ MinClassSize then
23:            DiscoveredSGroups ← DiscoveredSGroups ∪ {SGroup}
24:            SGroup ← φ
25:        end if
26:    end if
27: end for
28: return DiscoveredSGroups
```

Algorithm 6.1: Social-DBSCAN

6.5 Socialising Outside the Classroom

In the previous Section 6.4, we discussed how social density-based clustering of WLAN traces could be utilised to detect granular social groups of mobile users within a university campus. In particular, we proposed Social-DBSCAN to automatically detect the regular learning activities taking place at chosen locations, e.g. classrooms, and provide accurate estimates about the attendance levels. In this section, we study the social aspect of human presence outside the classroom, with the aim of gaining a better understanding of the presence and movements of students within the case-study environment - the Bloomsbury campus of Birkbeck, University of London. Motivated by the findings presented in the previous section, i.e. Section 6.4, and based on the evident temporal regularity in pattern of revisits seen in Figure 6.3, we can deduce that the visitors have preferences with respect to the visited locations as well as the times they visit these locations. Although most of these preferred locations are related to teaching, there are also those that are utilised for non-teaching purposes such as the Coffee-shop at the Malet Street site. As shown in Figure 6.3,

```
1: ExpandSocialGroup(p, SGroup, pSimilarPts, CohCoff, DiscoveredSGroups)
2: SGroup ← SGroup ∪ {p}
3: for each (q ∈ pSimilarPts) do
4:     if q ∉ VisitedPts then
5:         VisitedPts ← VisitedPts ∪ {q}
6:         set qSimilarPts ← FindSimilarPts(q, CohCoff)
7:         if |qSimilarPts| > 0 then
8:             # add all points in qSimilarPts to pSimilarPts.
9:             pSimilarPts ← pSimilarPts ∪ qSimilarPts
10:            SGroup ← SGroup ∪ {q}
11:        end if
12:    end if
13: end for
```

Algorithm 6.2: Social group expansion in Social-DBSCAN

there is high temporal regularity of revisit, which gives a strong indication of the existence of an underlying *semantic* grouping of the visitors of these locations. The explanation of this semantic grouping is related to the teaching timetable, which not only dictates the time and location of the learning sessions but also dictates when the students can have their break-times, and thus directly influences their presence and movement behaviour on campus. As discussed later in this section, we find that a substantial number of students, who attend the learning sessions, visit locations such as the Coffee-shop during break-times. We conjecture that similar patterns are likely to occur in other environments, such as the work place, where people visit similar kinds of locations during their lunch break.

6.5.1 Methodology

In this section, we are interested in identifying those groups of students that maintain a social connection outside the classroom. The basic idea for detecting such social groups can be summarised as follows:

Step 1. We utilise Social-DBSCAN to identify the group of students attending the same regular class. The details of how this task is performed are presented in Section 6.4.

Step 2. Parallel to Step 1 above, for every selected non-teaching location, e.g. the Coffee-shop, we apply the temporally-restricted version of Social-DBSCAN to cluster the students based on selected time intervals, comprising break-times. These time intervals are carefully chosen so that the detected group of students form *a cohesive social group*, i.e. a group of friends who are visiting the target location to socialise as opposed to a random group of students who, by chance,

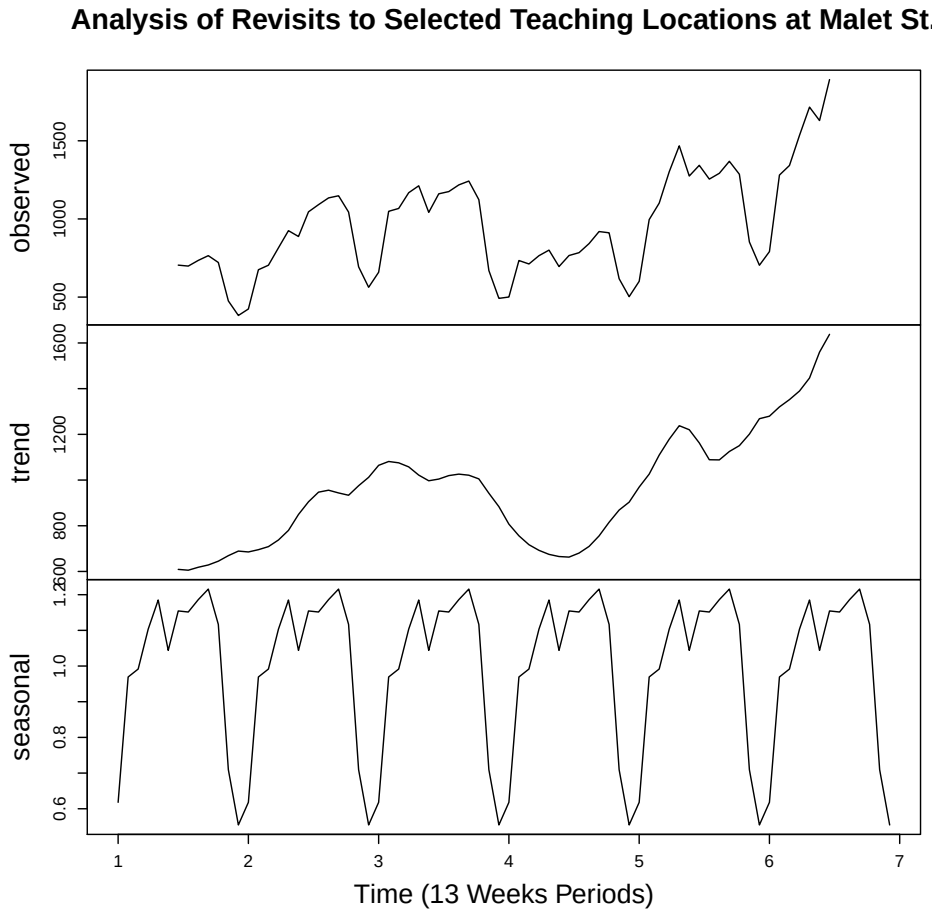


Figure 6.3: Time series analysis of number of revisits to selected teaching locations at the Malet Street site (see Table 6.3 for more information about these locations). In this figure, the top plot shows the original time series in which the data is divided into 13 week periods (Each 13 week period covering an 11 weeks academic term plus an extra week on either side of the term). The middle plot shows the estimated trend, and the bottom plot shows the estimated seasonal constituent [52]

happen to be in the same location at the same time. For the details of how this task is performed, the reader is referred to Section 6.5.4.

Step 3. We match the groups obtained in Step 1 to those obtained in Step 2, i.e. we identify the subsets of students, which contain those who attended non-teaching locations during the break times. The intuition here is that if a discovered cluster contains a group of users who attended the same class, then such a cluster consists of a group of friends. We provide a detailed description of how this task is carried out in Section 6.5.5.

6.5.2 Problem Formulation

Let S be a group of students, where each student is represented by his/her record of revisit to a target location L. Furthermore, assume that all the students in S visited the location L within a fixed time interval on a given weekday D for k consecutive weeks. We would like to perform the following tasks.

1. Automatically detect whether the collection of revisits, given in S, represent the attendance of a regular learning activity (i.e. class in this context) that was taking place at location L over the k consecutive weeks.

2. Detect whether the group discovered in the previous step contains subgroups of students that socialise at a target non-teaching location (e.g. the Coffee-shop).

In the remainder of this section we mention users, people, students and visitors interchangeably.

6.5.3 Detecting Regular Classes

In order to detect regular classes, we employ the Social-DBSCAN algorithm (see § 6.1) to discover coherent groups of students that attended the weekly sessions. As discussed in Section 6.4, the Social-DBSCAN is utilised in clustering the records of visits to the locations, where the learning sessions were taking place. Based on information from the semantic context, in which the algorithm is being applied, we determine the values of the parameters such as the *coherence of attendance* and *the minimum attendance threshold* values. Social-DBSCAN utilises the Jaccard coefficient to measure the distance between data points. The Jaccard distance measure plays a key role in capturing the similarity between the records of attendance of two students, i.e. it compares the total of shared attended sessions with the total number of sessions attended by either of the two students.

Following the clustering, we verify whether the presence of a detected group of students represents the attendance of a regular class that was taking place at the chosen location. The intuition for the clustering performed in Social-DBSCAN is that the students who regularly attended the location where the class sessions took place, should naturally form a single coherent social group.

From a practical perspective, there are two conditions that the clustering result must fulfil in order to verify the occurrence of a regular class at the chosen location.

1. A dominant group, with a significantly large number of the students (e.g. 50% or more) being members, must be one of the discovered groups.

2. The average number of students per session must be within the capacity of the chosen location, where the detected class sessions took place.

For further details of how the detection of regular classes is performed and how the values of the parameters of the Social-DBSCAN algorithm are determined, the reader is referred to the Section 6.4.

6.5.4 Socialising During Break-times

Using information extracted from the underlying academic context, we inform our model about the size of the time intervals that define the break-times. For example, we employ information extracted from the timetable and also the teaching practices followed at Birkbeck, where a three-hour lecture normally has two sessions divided by a 15 to 30 minutes coffee break. However, the time the break takes place is not fixed, and thus to detect the visit to locations such as the Coffee-shop, we undertake the following procedure.

Across the academic term, for each day of the week we make use of the records of visits to a target location such as the Coffee-shop, to compute time slots, of a maximum duration of n minutes, using the following sliding window technique.

Let W be the records of visits compiled for a given day of week over the 11 weeks academic term. To partition the set of records W, we assume a window of size n minutes. We extract from W the students records with visits that took place within the first n minutes, immediately after the end of the first hour of the regular teaching session. We then cluster these records to discover the social groups that met during these first n minutes. Next we then slide the window to include the minutes from the $(1 + k)th$ to $(n + k)th$ interval and attempt to to discover the social groups that met during this period. Consequently, we slide the window to include the minutes from the $(1 + 2k)th$ to $(n + 2k)th$ and so on until $n + mk$ is larger than a threshold value σ. Here, k is the number of minutes denoting the size of the window slide, i.e. number of minutes used for advancing the window along the time-line for the next iteration. We refer to the time interval of the sliding window, as the search window noting that the maximum threshold we used was 30 minutes.

6.5.4.1 Social Cohesion

While a *coherent* cluster of students attending the same class is defined as a group of individual users that have *similar* attendance over an 11 week academic term, a *cohesive social group* is a subset of a coherent cluster with *socially connected members* outside the classroom. In order to identify such a cohesive social group we introduce two extra constraints to the clustering process performed in Social-DBSCAN, which we described in Section 6.4. Firstly, in order to mitigate the situations where a social group comprises members representing a chain of *friend-of-a-friend*, which is a natural result of the pairwise clustering performed in Social-DBSCAN, we ensure that all the members of a discovered group are directly related to one another. Practically, a temporal constraint is imposed to ensure that the clustered individuals actually attended the observed location within a given time interval. More specifically, when we identify a social visit at an observed location, we restrict the maximum difference between the times of visits made, by the members of the group, to a fixed user-defined threshold value; for instance, the difference between the times of visits made, by two distinct members, is less than or equal to 5 minutes. Secondly, to ensure that the members of a detected group belong to a genuine social group we restrict the minimum number of meetings at the observed location to a user-defined threshold value (e.g. three meetings). The intuition here is that the more the members of the group meet the more likely that they belong to a genuine social group.

Similar to the Social-DBSCAN algorithm described in Section 6.4, the proposed clustering algorithm, which we call *Temporally-Restricted-Social-DBSCAN*, can be described in pseudo-code as shown in § 6.3.

The parameters *CohCoff*[1] and *MinSGroupSize*[2] (see code provided in § 6.3), represent the *Coherence Coefficient* and the *Minimum Size of a Social Group* threshold values, respectively. The parameter *MinNumShrdVists* represents the minimum number of shared visits (or meetings) that were made, to the observed location, by the detected group. Note here that the default value for *MinSGroupSize* is 2, as every member of a detected group must have at least one *friend*. The parameter *CohCoff* denotes the maximum value for the distance, i.e. $1 - Jaccard$, between the members of the same group. In order to identify a meeting between two or more students, their times of visit to the target location must fall within a given interval. The parameter *meetingSrtTime* and *meetingEndTime* denote the beginning and the end of such time interval. From a practical point of view, the values of all these parameters are highly dependent on the context in which the algorithm is being

[1] Coherence Coefficient
[2] Minimum Size of a Social Group

```
1: Temporally-Restricted-Social-DBSCAN(Dataset, CohCoff, MinSGroupSize, Min-
   NumShrdVists, n, k, sigma)
2: set DiscoveredSGroups ← φ
3: set SGroup ← φ
4: set VisitedPts ← φ
5: set NOISE ← φ
6: for each (p ∈ Dataset) do
7:     set i ← 0
8:     while (n + ik < sigma) do
9:         set meetingSrtTime ← 1 + ik
10:        set meetingEndTime ← n + ik
11:        if p ∈ VisitedPts then
12:            continue
13:        end if
14:        VisitedPts ← VisitedPts ∪ {p}
15:        # findSimilarPts() returns all points similar to p that
16:        # findSimilarPtshave a Jaccard distance ≤ CohCoff.
17:        pSimilarPts ← FindSimilarPts(P, CohCoff, MinNumShrdVists,
18:        meetingSrtTime, meetingEndTime)
19:        if |pSimilarPts| < MinSGroupSize then
20:            NOISE ← NOISE ∪ {P}
21:        else
22:            SGroup ← ExpandSocialGroup(p, SGroup, pSimilarPts,
23:            CohCoff, MinSGroupSize, MinNumShrdVists,
24:            meetingSrtTime, meetingEndTime)
25:            if SGroup ≠ φ then
26:                DiscoveredSGroups ← DiscoveredSGroups ∪ {SGroup}
27:                SGroup ← φ
28:            end if
29:        end if
30:        i ← i + 1
31:    end while
32: end for
33: return DiscoveredSGroups
```

Algorithm 6.3: Temporally-Restricted-Social-DBSCAN

applied.

For the pseudo-code illustrating how a discovered group can be expanded, the reader is referred to the pseudo-code provided in § 6.2.

6.5.4.2 *Temporally-Restricted-Social-DBSCAN* vs *Social-DBSCAN*

The key difference between these two algorithms, *Temporally-Restricted-Social-DBSCAN* and *Social-DBSCAN*, lies in the way the distance between two users is computed. Although both algorithms employ the Jaccard distance, nonetheless the distance is tightly related to the underpinning semantic context in which the algorithm is being applied -

Social-DBSCAN is utilised for learning activity detection whereas *Temporally-Restricted-Social-DBSCAN* is employed for social groups detection at a non-teaching location such as the Coffee-shop. While the duration of a targeted learning activity is fixed, e.g. between 18:00 to 21:00, hence all users being clustered are associated with a fixed time interval, visiting a non-teaching location is not restricted to a given time. Furthermore, the distance computation performed in *Social-DBSCAN* is only concerned with what sessions were attended by the two compared users and whether there were any matching sessions, irrespective of the time an observed user visited the location during any of the class sessions - effectively only a single dimension of the data is considered in the computation of the distance between two observed users. In contrast, the distance computation in *Temporally-Restricted-Social-DBSCAN* takes into consideration the time an observed location was visited by each user - *in order for two observed users to share the same cluster, the difference between their times of visit must be within a chosen threshold.*

A further distinction between the two algorithms is that *Temporally-Restricted-Social-DBSCAN* does not require the users being clustered to be given in any particular order. *Social-DBSCAN* however, utilises the level of attendance in detecting the group of students that attend an observed learning activity - it relies on the ordering of the users based on attendance level, where the higher the attendance the more likely that the user being clustered is part of the group that attended the regular class.

Due to the computation of the neighbourhood for each point in the data set, the two algorithms have similar complexity, where the worst-case complexity for both algorithms is $O(n^2)$.

6.5.5 Discovering Classroom-based Social Groups

Following the detection of regular classes, we deploy the *Temporally-Restricted-Social-DBSCAN* to discover the social groups, which regularly visited the Coffee-shop, from each class. Using a given time window length, *Temporally-Restricted-Social-DBSCAN* compares the records of visits of different individuals in order to detect shared visits, i.e. visits made to the Coffee-shop at approximately the same time. Consequently, those individuals that have shared visits to the Coffee-shop and attended the same classes are clustered into classroom-based social groups.

Property	Eduroam data set
Number of users	167,272
Average number of user-days	15
Average number of locations per user	10

Table 6.2: Properties of the Eduroam data set (for more details see Subsection 1.6.1.2).

6.6 Evaluation

6.6.1 Data Set

The evaluation of the proposed Social-DBSCAN and Temporally-Restricted-Social-DBSCAN is based on recent WLAN traces collected at Birkbeck - the case-study university in this research work. A detailed description of the data set has been provided in the evaluation section in the previous chapter (5.6.1). A summary of its properties is presented in Table 6.2.

In order to detect the attendance of classes, we create spatio-temporal vectors, where each vector denotes the visits made, by one of the users, at a given spatio-temporal context (i.e. visits made to a specific target location, within a fixed time interval of day, on a target weekday and over a period of 11 weeks). The data division into 11 weeks periods is motivated by the temporal regularity found in the data as shown in Figure 6.3, where each 11 week period corresponds to a single academic term. Also, in this chapter, we utilise the room capacity information, which is available independently through Birkbeck's website. All of the 15 locations we chose for the evaluation of the proposed clustering approach are rooms with known student capacities.

6.6.2 Experiments

6.6.2.1 Detecting the Attendance of Learning Activities

Our evaluation methodology is designed to verify that the presence of a discovered group of individuals represents a regular attendance of learning activity, which has been taking place at the observed location. It compares the results obtained from the two stage process, which addresses the noise reduction and the coherence of attendance, against the initial intuition that the regular visitors of an observed location on a given day of the week is mainly comprised of a single coherent social group (see the discussion in Section 6.4.3 for more details). From a practical point of view, there are two criteria that the clustering result must fulfil in order to justify the occurrence of a regular class at the target location:

1. There must be a dominant *coherent* discovered group with the majority of the students being members of such a group.

2. Following the reduction of noise, the average number of students per session must be within the capacity of the target location, where the detected class was taking place.

6.6.2.2 Detecting Classroom-based Social Groups

Our testing methodology herein is designed to verify the presence of groups that socialise at a non-teaching location such as the Coffee-shop. The distinctive property of the target social groups is that each comprises individuals who attended the same classes. The experiment involves carrying out the following sequence of steps:

Step 1. Detect whether the set of revisits, recorded at a selected location, represent the attendance of a regular learning activity (i.e. class in this context) that was taking place at the chosen location over the observed consecutive weeks.

Step 2. Find out whether the set of individuals, who attended the class, detected in Step 1, contains subgroups of students that socialise at one of the non-teaching locations (e.g. the Coffee-shop).

6.6.3 Results

6.6.3.1 Attending Learning Activities

In our experiments, we examined the visiting patterns from 15 chosen locations with known capacity. These chosen locations are usually used for learning activities such as lectures and lab-based classes. As shown in Table 6.3, the number of unique visitors greatly varies between these chosen locations. For a few of them, the number of visitors exceeds the location capacity, which is a clear indication that those visitors were not all regular class attendees at these locations. Therefore, it is important that we remove the noise from the data and only preserve the records of those visitors who most likely visited those locations in order to attend the learning activities that were taking place there. In order to ensure that only individuals with consistent high attendance are being clustered, we decided to filter out the records of those individuals with attendance less than 40%, i.e. students with attendance of four sessions or more. The intuition here is that the group of those individuals with regular attendance of 40% or more would include all those who attend the actual class and most probably exclude all the *noise*, i.e. individuals who are not

regular attendees of the class such as those who happened to be in the vicinity of the target location when the class was taking place.

We restricted our investigation to the data covering the Spring of 2015 (i.e. the period from 5th Jan - 20th March 2015). The results shown are for the time interval from 18:00 - 21:00 of every Monday of this period. The obtained results are shown in Table 6.4, and are also summarised in the following two points.

1. With the exception of location #1, for every location, the average number of students per session was always smaller than the capacity of the target location. After the verification against the timetable, it appears that location #1 hosted two classes on Monday evening; one class running from 18:00 - 19:30 and the other from 19:30 - 21:00. The fact that the average number of students per session for the location was only 11% less than twice the capacity, confirms that the clustering result is consistent with the finding that the location hosted two sessions on Monday evening.

2. As shown in Table 6.4, the average number of students per session for some of the locations was far too small in comparison to the capacity of the target location (e.g. locations #14 and #15). Such situation can be attributed to the possibility that a substantiated number of students in those classes might not have been active Eduroam users.

6.6.3.2 Detecting Social Activities Outside the Classroom

Recorded Number of Visitors

By deploying the *Temporally-Restricted-Social-DBSCAN* using the records of visits made to the Coffee-shop, we can detect the social groups of students who attend the same classes. The result summarised in Table 6.5, shows the hourly social activity at the Coffee-shop during the 11 weeks period of the Spring term in 2015. In this result a *Coherence Coefficient* value of 0.7 was utilised and each pair of students in a social group, shared at least two meetings at the Coffee-shop.

As shown in the Table 6.5, for some of the time intervals, the number of visitors are significantly larger in comparison to others. This is most likely because the visitors to the Coffee-shop at these times represent a number of individuals that we did not include in our study, e.g. students attending classes that we did not include in our experiment.

Location ID.	Site	Location Name	Number of Unique Visitors	Capacity
#1	MaletSt-402	Malet Steet	239	35
#2	MaletSt-G16	Malet Steet	194	60
#3	Clore-102	Clore Management Centre	135	33
#4	MaletSt-b35	Malet Steet	72	125
#5	MaletSt-153	Malet Steet	66	66
#6	MaletSt-b34	Malet Steet	43	222
#7	MaletSt-b29	Malet Steet	48	30
#8	MaletSt-417	Malet Steet	49	60
#9	MaletSt-423	Malet Steet	42	39
#10	Clore-204	Clore Management Centre	32	33
#11	MaletSt-352	Malet Steet	21	20
#12	43GordonSq-g02	43 Gordon Square	24	28
#13	MaletSt-314	Malet Steet	32	36
#14	MaletSt-b20	Malet Steet	16	99
#15	43GordonSq-b04	43 Gordon Square	17	127

Table 6.3: Location information.

Location ID.	Number of Students	Number of Discovered Groups	Group No.	Group Size	Group Attendance			Standard Deviation	Avg. Number of Students Per Session
					Min	Max	Avg.		
#1	118	1	1	116	5	10	5.94	1.08	62.64
#2	102	1	1	99	5	9	6.00	1.15	54.00
#3	62	3	1	47	5	10	6.28	1.22	31.91
	62	3	2	8	5	6	5.13	0.33	
	62	3	3	3	5	5	5.00	0.00	
#4	53	1	1	50	5	10	6.40	1.40	29.09
#5	34	1	1	28	5	10	6.07	1.46	15.45
#6	32	1	1	30	5	9	6.93	1.26	18.91
#7	27	3	1	13	5	9	7.62	1.21	12.18
	27	3	2	4	6	7	6.25	0.43	
	27	3	3	2	5	5	5.00	0.00	
#8	32	2	1	27	5	9	7.04	1.23	18.18
	32	2	2	2	5	5	5.00	0.00	
#9	26	3	1	15	5	8	6.87	1.02	11.18
	26	3	2	2	5	5	5.00	0.00	
	26	3	3	2	5	5	5.00	0.00	
#10	17	3	1	9	5	9	7.11	1.20	7.64
	17	3	2	2	5	5	5.00	0.00	
	17	3	3	2	5	5	5.00	0.00	
#11	17	2	1	8	5	10	6.75	1.71	6.27
	17	2	2	3	5	5	5.00	0.00	
#12	15	1	1	13	5	9	6.46	1.34	7.64
#13	16	2	1	9	5	9	6.67	1.33	6.55
	16	2	2	2	6	6	6.00	0.00	
#14	10	1	1	6	5	8	6.00	1.15	3.27
#15	11	2	1	2	9	10	9.50	0.50	5.45
	11	2	2	7	5	7	5.86	0.99	

Table 6.4: Social-DBSCAN clustering result for 15 unique locations. The student's minimum attendance threshold was 40% and the Coherence Coefficient (*CohCoff*) was 0.6. This result was computed for the time interval from 18:00 - 21:00 every Monday of the Spring term of 2015 (11 weeks period).

Time Interval	Number of Visitors	Number of Classes	Number of Visiting Groups	Average Group Size	Group Min Attendance	Group Max Attendance	Group Avg Attendance	Standard Deviation
0:00-1:00	86	1	1	1.0000	3	1	1.0000	0
7:00-8:00	139	1	1	1.0000	3	1	1.0000	0
8:00-9:00	1767	6	8	1.5000	3	5	2.3750	1.44
9:00-10:00	4327	6	10	1.6000	3	10	4.9000	3.91
10:00-11:00	6077	7	12	1.4167	3	11	5.0000	4.77
11:00-12:00	6508	7	15	1.4667	3	12	4.4000	5.21
12:00-13:00	7208	8	16	1.4375	3	11	4.3125	5.05
13:00-14:00	8541	6	16	1.6250	3	12	5.6875	6.33
14:00-15:00	6403	9	22	1.5455	3	12	4.4091	6.82
15:00-16:00	6868	8	22	1.6364	3	12	5.0000	6.97
16:00-17:00	6494	8	22	1.7273	3	12	5.0455	8.09
17:00-18:00	6546	9	30	2.2000	4	14	7.4000	11.50
18:00-19:00	4117	8	29	2.0690	3	11	5.7586	8.66
19:00-20:00	3689	9	29	2.1724	3	22	7.6207	11.49
20:00-21:00	3081	8	28	2.0714	4	13	6.3214	10.17
21:00-22:00	2124	8	26	1.8846	3	9	3.9231	6.46
22:00-23:00	1138	4	9	1.4444	3	7	2.8889	3.25
23:00-0:00	780	7	12	1.3333	3	8	2.5000	3.13

Table 6.5: Detected Social Activity at the Coffeeshop

Moreover, the Coffee-shop is frequently visited by large numbers of people, who are not students of Birkbeck, but make a considerable proportion of the total number of visitors of the Coffee-shop.

Detected Social Groups

As shown in Figure 6.4, the number of social groups, which consistently met at the Coffee-shop is small in comparison to the total number of those who attended the classes and relative to the number of social groups, given in Table 6.5. Nonetheless, as a proof of concept, the results are promising and more experiments need to be carried out on another more complete data set, as opposed to the one used in this analysis. A richer data set in which the users authentication information is recorded more frequently would certainly allow for more extensive analysis to be carried out.

It is interesting to discover that a significantly large social activity occurs at the break halfway through the three hour evening lecture. As shown in Figure 6.5, a total of 94 groups met at the Coffee-shop during the time from 19:15 to 19:45. It is equally interesting to find out that the largest social activity occurred during the time from 17:45 to 18:00, just before the start of the class at 18:00.

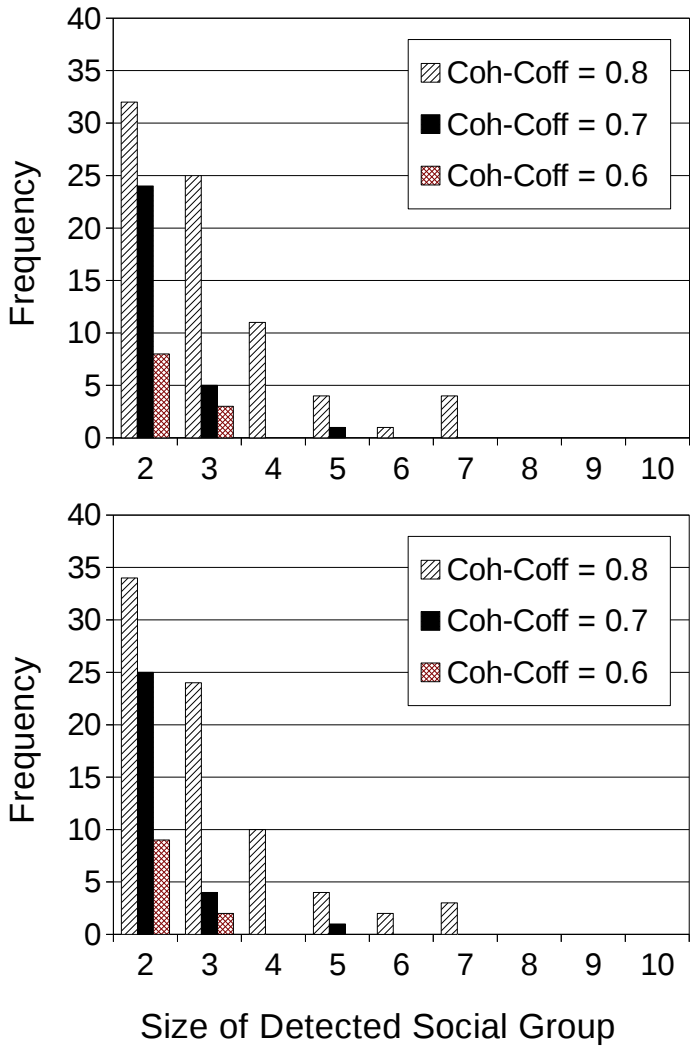


Figure 6.4: Distributions of the size of detected social groups. Each distribution is denoted by a different colour and computed for a given Coherence Coefficient value. In the top figure, each pair of students in a social group, shared at least three meetings at the Coffee-shop. In the bottom figure, each group had at least four meetings.

6.7 Discussion

6.7.1 Demoting Users for Poor Attendance

When discovering a class, it is usual to have noise due to the Wi-Fi detecting all movement within the vicinity of a target location. One way to ensure that such noise is filtered out from the data is to approximate the time spent at the target location and discount those individuals who spent a short time in the vicinity of the location. However, an individual who is present in the vicinity with no intention to attend the class at the target location is

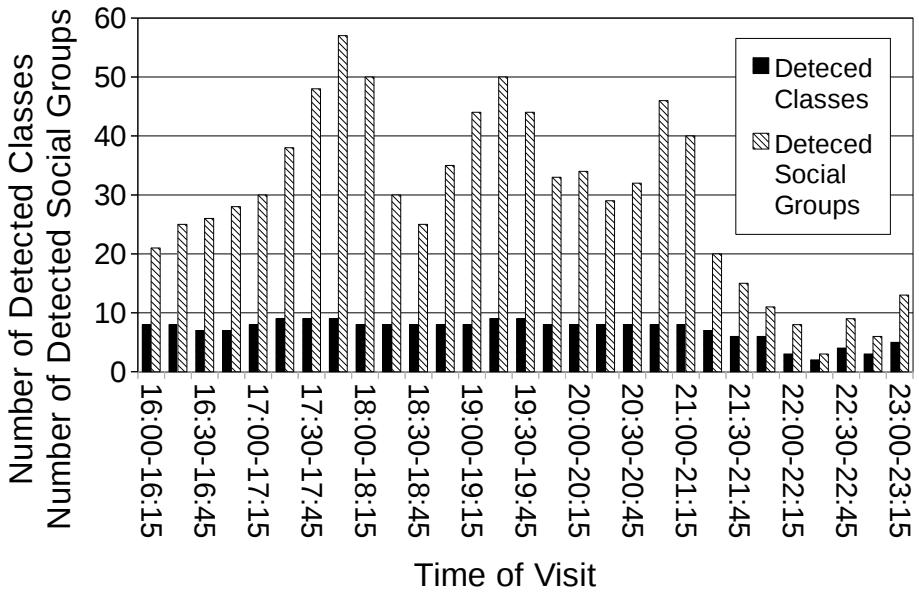


Figure 6.5: Distribution of number of detected social groups and the number of classes, to which the detected groups belong, by time of day. These distributions were computed using a Coherence Coefficient value of 0.7, where the members of each detected social group had at least two meetings at the Coffee-shop.

very unlikely to consistently appear at the same location at the same time of the day over the 11 week period. Thus, the noise due to such inconsistent appearance at the target location can be removed by raising the minimum attendance threshold, which ensures that only individuals with consistent attendance remain in the data.

6.7.2 Robustness Against Incoherent Revisits

One of the very attractive features that our proposed Social-DBSCAN shares with DBSCAN is the robustness to outliers [67] and [35], which Social-DBSCAN capitalises on to ensure that the discovered groups do not contain any incoherent member points. However, even after filtering noise out and clustering the points by using the Jaccard-based distance, we may still discover more than one coherent group. In the context of class detection, the occurrence of such a scenario can be attributed to the possibility that there may be two classes sharing the period from 18:00-21:00, e.g. one class running from 18:00 - 19:30 and another running from 19:30 - 21:00.

Another explanation for discovering more than one group is based on the fact that some students may have irregular attendance patterns. The dataset used for the evaluation contains records of attendance of many mature students, who have daytime jobs

and may occasionally miss classes due to some special circumstances such as unexpected additional work commitment. In such cases, the majority of the students are usually clustered together in one single large group while those students with irregular attendance form a small-sized group or groups (see the result for location #3 in Table 6.4). In any case, Social-DBSCAN ensures that incoherent behaviour is separated from the dominant coherent pattern extracted from the data.

6.7.3 Border Points

Similar to DBSCAN, Social-DBSCAN cannot determine the clusters of border points that are reachable from more than one cluster. For example, a student that is equally part of two different learning sessions where s/he some time arrives four times just before the end of the first session and four other times after the session has ended. Such a student can be considered a late attendee of the first session based on those occasions in which s/he arrived before the end of the session. However, he can also be an early attendee of the second session based on those occasions in which he arrived late for the first session and early for the second. In the context of class detection, we overcome such a problem by restricting the the detection of an observed class attendance to a given time interval, for instance, 18:00 - 19:30, where any arrival after 19:30 would not be considered.

6.7.4 Estimating Class Attendance

By combining external data from the teaching timetable with the social, spatial and temporal information extracted from the dataset, we can identify groups of users whose presence at a given spatio-temporal context can be linked with the learning activity taking place within the same context. Moreover, the size of the discovered group of students can potentially give an accurate estimate of the class attendance.

We acknowledge herein that a substantial proportion of the students who attend classes are not active users of Eduroam. However, Eduroam is increasingly becoming more pervasive and more popular amongst the regular and non-regular visitors of Birkbeck. The analysis of more recently recorded traces shows that there is a steady increase in the number of Eduroam users across the university's main site.

6.7.5 Sensitivity to Time of Visit

A key advantage of the proposed Temporally-Restricted-Social-DBSCAN, over the original Social-DBSCAN is that the proposed algorithm is capable of clustering visitors in terms of

time of visit to a target location, in order to discover social connections. Utilising a fixed size time intervals, the algorithm can successfully identify the visits made to the Coffee-shop at approximately the same time - it can successfully detect the meetings of different individuals at an observed socialising location. This allows the *algorithm* to compare the different records of visits, and in turn, to cluster visitors into social groups.

6.7.6 Lack of Proximity Data

The ability to measure the proximity between *co-located* individuals, during a visit to an observed location, is a key factor in accurately inferring whether the individuals are socialising or visiting the target location for different reasons - for example when two students visit the Coffee-shop after the class, but sit at separate tables. Unfortunately, the data set we utilised for the evaluation of the proposed methods does not contain any between visitors' proximity information. Evaluating the proposed methods using a richer data set that provides information about the proximity between users visiting a target location will most likely increase the accuracy of the obtained results.

6.8 Summary

1. We showed how by clustering WLAN activity traces we can detect granular social groups of mobile users within a target academic environment. Moreover, we showed how by being able to detect social groups at target locations, we provide an invaluable opportunity to understand the presence and movement of people within such an environment. Using the proposed Social-DBSCAN, we demonstrated how we can automatically detect the regular classes taking place at target locations, and provide accurate estimates about the levels of attendance.

2. We illustrated that by using the proposed methods, namely Temporally-Restricted-Social-DBSCAN and Social-DBSCAN, we can automatically detect regular learning activities, and discover social groups among the students, who attend these activities.

Chapter 7

Formal and Informal Social Spaces

7.1 Overview

The different kinds of activities that take place within an observed learning environment such as a university campus determine to a large extent the kind of social interactions exhibited by the users in such environments. Using a big data set of Wi-Fi activity traces, we attempt, in this chapter, to understand how these social interactions characterise the space within an observed university campus. We discovered that there are at least two types of social interactions within a university campus: *formal* such as attending a class and *informal* such as meeting friends at the cafeteria. Each of these two types of social interactions is associated with a specific set of locations within the observed campus. We also discovered that users tend to restrict their social interactions to a small set of geographical locations, and often revisit the same location to socialise with the same social group. Also, irrespective of the type of the social interactions, users tend to restrict their revisits to geographically nearby locations and only revisit locations that are further afield when they are in the company of their social group. These findings are based on the social groups detected by a scalable density-based clustering method applied to a large data set of mobile users Wi-Fi traces. The results of the experiments carried out in this research demonstrate how the proposed algorithm (see Algorithm 7.1 in § 7.5.2) can non-invasively detect social groups on the basis of the activity performed at the observed location.

7.2 Introduction

The detailed information produced by Wi-Fi provides an invaluable opportunity to learn about the different aspects of presence and movement behaviours of people within a given

environment such as an organisation office complex or a university campus. With the aid of appropriate tools, for analysing such detailed information, we can potentially discover hidden patterns of behaviour at both the collective and the individual user levels; thus we increase our understanding about people's presence, and, in turn, improve our ability to make informed decisions when we plan for our environments.

This chapter is organised as follows. In Section 7.3, we describe the primary objective and the contributions of the research work presented herein. Section 7.4 discusses the characteristics of social activities within the case-study environment, particularly those that occur at informal social locations such as the Coffee-shop. In Section 7.5, we investigate the different types of social presence across the university campus and present the proposed algorithm (see Algorithm 7.1 in § 7.5.2) for detecting the social grouping of users. Section 7.6, presents a model for formal and informal locations based on the social interactions that take place across the case study environment. In the Evaluation section, i.e. Section 7.7, we give a description of the data set used for the evaluation of our proposed approaches, and provide a comprehensive discussion of the results of the experiments. Finally, in Section 7.8, we provide a brief discussion about the lack of proximity information in the evaluation data set, and we conclude the chapter with a summary statement, which can be found in Section 7.9.

7.3 Problem Definition

The key research objective of this chapter is to characterise the space within an observed university campus. More specifically, we would like to determine the type of an occupied location based on the visiting behaviour exhibited by the social groups that visit such a location. We first focus on the detection of granular social groups of mobile users, within the university campus, on the basis of the social activities that take place at observed locations. The intuition herein is that the social activities in which the detected groups participate can be categorised into *formal* and *informal* activities where each category is associated with a set of specific locations within the university campus.

7.3.1 Contributions

This chapter makes the following contributions:

1. It proposes a density-based clustering method that discovers social groups by utilising activity traces of mobile users. We detect the social groups on the basis of the

activities taking place at observed locations within a university campus. We provide a detailed description of this clustering method in Section 7.6.

2. It proposes a framework for inferring the type of an observed location, by using the patterns of visit extracted from Wi-Fi activity traces. Here we have two main types of social locations: *formal* and *informal*, which we define in the next section, namely Subsection 7.5.1.

3. It investigates the similarities and differences between the *formal* and the *informal* social locations.

7.4 Characteristics of Social Spaces

Understanding the social dynamics within an observed environment such as a university campus can be useful for a range of applications. In this chapter, we study the social aspect of human presence with the aim of gaining a better understanding of the presence and movements of people within the case-study environment - the Bloomsbury campus of Birkbeck, University of London. By knowing *who* and *where* and *why* people spend their time, the university can plan for the most effective usage of space and allocation of services in the manner that creates a more positive attitude toward learning, and provides a richer and more rewarding experience.

7.4.1 Types of Social Behaviour

The numerous daily activities that take place at the case-study environment, which include "learning classes", "meetings", "seminars" and "having lunch at the cafeteria", can be broadly divided into two main categories: *formal* and *informal* activities. Generally, in a *formal activity*, such as a learning class or a seminar, the social interaction is between a large group of individuals taking part in the activity, whereas in an *informal activity* we tend to find a close social interaction between a relatively smaller group of individuals. Moreover, individuals usually spend roughly the same duration of time when they attend a formal activity session whereas they tend to spend variable length of time when they are involved in an informal activity. Also, formal activities are usually linked to specific locations and appear to follow a regular pattern of occurrence whereas informal activities tend to not adhere to a fixed pattern of occurrence.

Generally, these two categories of activities underpin the different types of social behaviour that can be found at our chosen environment. In this chapter, we distinguish between two kinds of social presence: *formal* and *informal*, which we interpret as follows:

Formal Social Presence

Formal Social Presence denotes the set of meetings that are attended by the same group of individuals, take place at the same location and occur regularly in sessions of fixed duration. We refer to the type of social relationship exhibited in such set of meetings as *formal social relationship* and the social group of users, who participate in such a relationship, as *formal social group*.

Example: The regular meetings of a group of the same students attending a three-hour weekly lecture that takes place at a specific lecture-room at specific time, e.g. from 18:00 - 21:00 every Thursday of the Spring academic term.

Informal Social Presence

Informal social presence is defined as the set of meetings that are attended by the same group of individuals and may take place at different locations. In contrast to the meetings of the *Formal Social Presence*, these meetings do not necessarily follow regular patterns of occurrence or have a fixed duration. We refer to the type of social relationship shown in such meetings as *informal social relationship* and the social group of mobile users, who take part in them, as *informal social group*.

Example: The meetings of the same group of friends at a cafeteria for coffee. More often these meetings, attended by the same group of friends, take place at different locations, e.g. a different cafeteria or at the cinema. Moreover, these meetings are usually irregular in their occurrence, i.e. happen at different times or have different durations.

In this section, we use the term *'visit'* to refer to an event when the time and the location of a particular user is recorded. This means that a user was at a specific location (i.e. a room) when s/he either initiated or received data using their mobile device over Wi-Fi.

7.4.2 Types of Visited Locations

Unlike localisation techniques, which focus on discovering the exact location of the mobile device, in this section we are only interested in determining whether two, or more, devices

are within the same room. We selected two types of locations for the evaluation of our proposed method: meeting rooms, where regular learning and administrative activities take place and leisure locations with food and drinks facilities. The details of these locations are given in Table 7.1.

Location	Site	Category	Number of Visitors
Bar	Malet St. Ext.	Leisure (informal)	4677
Cinema	43 Gordon Sq.	Leisure (informal)	3035
CoffeeShop	43 Gordon Sq.	Leisure (informal)	2967
CoffeeShop	Malet St.	Leisure (informal)	38520
Room 102	10 Gower St.	Learning (formal)	9963
Room 301	Malet St.	Learning (formal)	4249
Room 314	Malet St.	Learning (formal)	7076
Room 413	Malet St.	Learning (formal)	665
Room 417	Malet St.	Learning (formal)	189
Room B29	Malet St.	Learning (formal)	16081
Room 254	Malet St. Ext.	Learning (formal)	19051
Room 456	Malet St. Ext.	Learning (formal)	12031

Table 7.1: Selected Birkbeck Locations

7.4.2.1 Patterns of Visits

We studied the number of revisits made to locations across the campus and we observed that the distributions follow a power law for most locations. Figure 7.1 plots the distributions for the number of revisits made to locations where *informal activities* occur: the Coffee Shop and the Cinema at 43 Gordon Square, the Bar and the Coffee Shop at Malet Street. The log-log plots in this figure unanimously show that the distributions follow a broken power law consisting of two power law regimes - the broken power law in this case is a function comprising two power law distributions combined with some threshold [64]; for example, with two power laws. Initially, for the first two revisits, the distributions climb to their peak points at slopes 3.51, 2.76, 3.62 and 4.72, respectively. Then, for up to 25, 31, 25 and 63 visits, they descend gently at slopes -2.58, -2.26, -1.87 and -2.29, respectively. The four distributions jitter sharply for values of revisits beyond these ranges, which is a sign of an exponential cutoff [40]. Interestingly, those individuals who made their first visit are *more likely* to revisit the observed location. This pattern suddenly reverses across

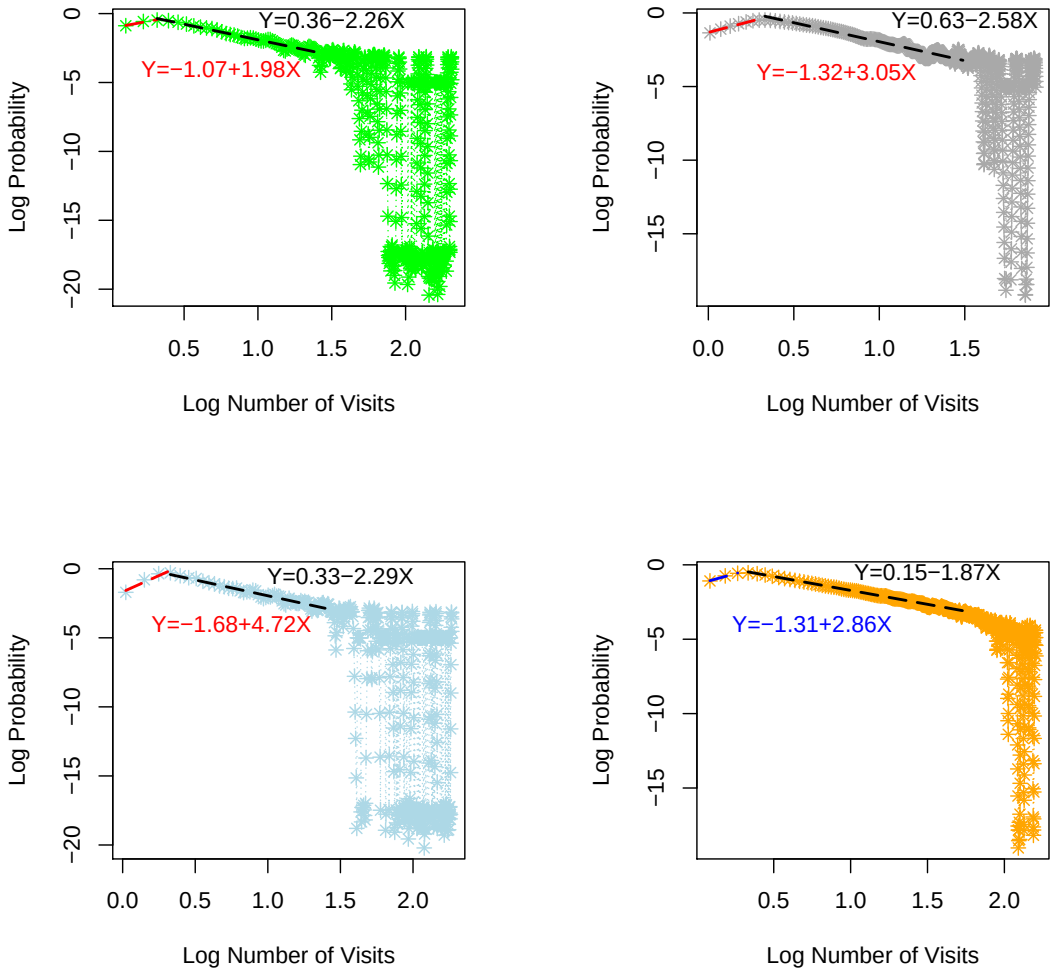


Figure 7.1: Distributions of number of revisits to the locations where *informal activities* occur. A revisiting user is one who made two or more visits to an observed location. Shown from left to right are the distributions for: the Coffee Shop, the Cinema at 43 Gordon Square, the Bar and the Coffee Shop at Malet Street. The two fitted straight lines indicate the broken power law relationship in each plot.

the four locations, where for those individuals who made between 3 to 25, 3 to 31, 3 to 25 and 3 to 63 visits, respectively, the higher the number of their previous visits the *less likely* that they will revisit the observed location.

7.5 Detecting Different Types of Social Presence

Our intuition is that the activity taking place at an observed location determines, to a large extent, the kind of social interaction that occurs during the activity. Methods that only capitalise on temporal and spatial information to detect social groups of people visiting an observed location, may not always produce the desired accurate results. For example, during a formal meeting or a seminar, people may be seated far from one another despite being closely related to each other. Equally, they may be seated adjacent to one another despite the lack of a close relationship between them. A method that solely depends on proximity information to detect the social group attending a meeting or a seminar in which individual people are placed at distances greater than what is required to link them to one another, will most probably fail to detect the correct social grouping. Similarly, a clustering method that relies on a small distance between arrival times, will fail to correctly cluster two individuals that attend a meeting but arrived at times far apart from one another. Equally, a method that expects individuals' arrival times to be long apart from one another, will fail to detect a social event that occurs within shorter *time intervals*. For example a method, designed to detect groups that attend social events in which individuals arrive an hour apart from one another, will most likely fail to discover short events such as a 15 minutes coffee-break gathering at the cafeteria. We argue here that in order to detect the correct social behaviour at a given location, it is imperative that, in addition to the temporal and spacial information, we take into consideration the semantic underpinning of the social interaction at that location. For example, a clustering method that adapts to different social activities will be able to adjust its temporal and spacial criteria in order to correctly detect the social group attending such meetings. Our proposed clustering method (see Algorithm 7.1 in § 7.5.2), which we discuss in the next subsection, i.e. Subsection 7.5.1, is parameterised with information about the kind of activity that takes place at an observed location.

7.5.1 Social Density-based Clustering

We adopt a density-based clustering style such as the one implemented in DBSCAN [35] for our proposed clustering algorithm. Our motivation for such a style stems from the fact that density-based algorithms are not restricted to discovering only clusters that are spherical in shape. These clustering algorithms can discover any arbitrary-shaped clusters which other alternatives such as the partitioning algorithms, e.g. K-means [110], and

hierarchical algorithms, e.g. divisive hierarchical clustering [106] would not be able to accurately identify - such alternative algorithms inaccurately identify convex-shaped regions, where outliers are usually present in the identified clusters.

Building on the previously mentioned intuition (see the start of Section 7.5), we propose a new scalable method that detects the social clustering of mobile users on the basis of the type of activity performed at an observed location. Given a database of users and a set of locations, we would like to discover the groups of users that visit these locations to participate in a social activity. For example, we would like to discover groups of students who attend lectures together as classes at different lecture-rooms, groups of researchers who hold regular seminars at particular meeting rooms or groups of friends who socialise at the Coffee Shop during break time.

In order to formulate how we would discover such social groups we would like to introduce the notation provided in Table 7.2.

Symbol	Meaning
U	Database of users.
L	The set of locations.
p, q, r	An m-dimensional point representing a user's set of visits to the locations given in L.
v	A user's visit, to a given location, within a time interval t.
D	The set of m-dimensional points representing the users in U.
$\theta_{q,r}$	The Jaccard distance between q and r, (see § 6.4.2.4)
$RN_\epsilon(p)$	The neighbourhood of p in which the maximum distance between any pair of points is ϵ.
G	A social group of users.
δ	The minimum number of joint visits
$minPts$	A density threshold.

Table 7.2: Notation

The core concept of the proposed SocialDBC algorithm (see Algorithm 7.1 in § 7.5.2) for social clustering is that a data point is assigned to a cluster/group if it is *socially-connected* to all the other member points of the cluster or the group. To explain this key idea, we give the following definitions of concepts that are common to many density-based clustering algorithms such as DBSCAN [35].

Given a data set of points D, SocialDBC estimates the density around p using the concept of ϵ-restricted-neighbourhood, which is defined as hereafter.

Definition 1. ϵ-neighbourhood

An ϵ-neighbourhood, $N_\epsilon(p)$, is the spherical space with radius ϵ and p at its centre. This is formally defined as follows:

$$N_\epsilon(p) = \{r \mid \theta_{p,r} \leq \epsilon\}. \tag{7.1}$$

Definition 2. ϵ-restricted-neighbourhood

An ϵ-restricted-neighbourhood, $RN_\epsilon(p)$, is the neighbourhood of p in which the maximum distance between any pair of points is ϵ. This is formally defined as follows:

$$RN_\epsilon(p) = \{q \mid q, r \in N_\epsilon(p),\ \theta_{q,r} \leq \epsilon\}. \tag{7.2}$$

Note that the point p is always a member of its own ϵ-restricted-neighbourhood, i.e., $p \in RN_\epsilon(p)$ always holds.

Given the above definition, one can see that the neighbourhood $RN_\epsilon(p)$ is a subset of the ϵ-neighbourhood $N_\epsilon(p)$, in which any pair of points are within a maximum distance ϵ, thus

$$|RN_\epsilon(p)| \leq |N_\epsilon(p)| \tag{7.3}$$

Definition 3. Core Point

A point $p \in D$ is classified as:

1. a *core point* if its neighbourhood $N_\epsilon(p)$ has high density, i.e., $|N_\epsilon(p)| \geq minPts$, where $minPts$ is a user-specified finite positive integer,
2. a *noise point*, otherwise.

Definition 4. Social Connectivity

Two points p and $q \in D$ are *socially connected* if:

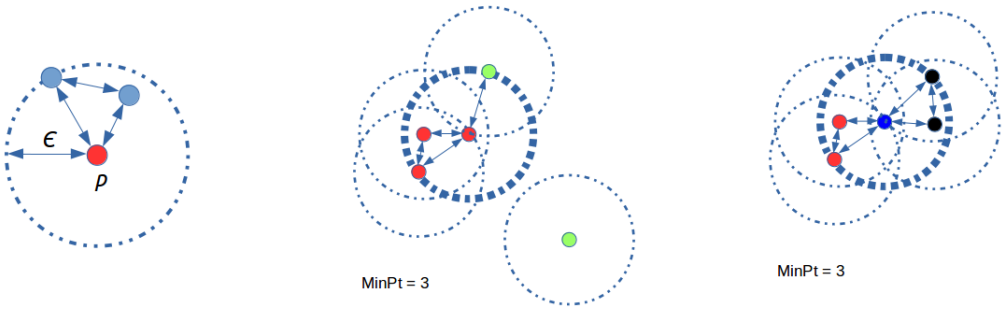
1. $q \in RN_\epsilon(p)$,
2. and the neighbourhood $RN_\epsilon(p)$ has high density, i.e., $|RN_\epsilon(p)| \geq \delta$, where δ is a user-specified finite positive integer denoting a minimum density threshold.

Definition 5. Social Group

A social group, G, is a *socially connected* set of points. An example of such a social group can be found in a group of students that attend the *same* weekly class learning sessions that are at least equal to δ sessions in total. Such a group is socially connected because every member of the group attended at least δ sessions that the other members attended irrespective of whether the sessions took place at one or several locations.

In order to make the above mentioned definitions clearer, we refer the reader to the example given in Figure 7.2, in which we visualise the concepts of ϵ-restricted-neighbourhood, core and noise points, and multi-cluster membership, respectively. In the subfigure (a), the red small circle represents a point *'p'*. The unlabelled double-sided arrow represents a distance that is less than or equal to ϵ, whereas the dotted circle denotes the area representing the neighbourhood of *'p'*. Within this neighbourhood, the two small circles filled with blue colour denote two random neighbouring points of *'p'*. The distinctive feature of all points in this neighbourhood of *'p'*, including *'p'* itself, is that they are located at a distance less than, or equal to, ϵ from each other. In the subfigure (b) the bold-dotted circle denotes a neighbourhood in which the three red points are at distance less than or equal to ϵ, and thus considered to be *core points*, whereas the point coloured in green is considered to be *noise* despite being at a distance less than or equal ϵ. This is because the green point is located in the ϵ-restricted-neighbourhood of only one of the red points and not all three red points; and thus falls short, by two points, in satisfying the minimum number of neighbouring points criterion for core points, i.e. the user-specified minimum density threshold (minPts), which is equal to 3 in this example. The subfigure (c) visualises the concept of multi-cluster (or multi-group) membership which is shown by two clusters sharing a common point between them; the first of these two clusters is denoted by the two red points and the second cluster is represented by the two black points. The third point in each of the two clusters is denoted by the same point which is filled with blue colour. Although the points in this subfigure fall inside the same bold-dotted circle, which has a radius less than or equal to ϵ, and has the shared blue point being at the

centre, they are not part of the same ϵ-restricted-neighbourhood. This is because neither of the two red points fall within a distance less than, or equal to, ϵ from any of the two black points - in this figure any distance between two points that is less than or equal to ϵ is denoted by double-sided arrow.



(a) (b) (c)

Figure 7.2: (a) The ϵ-restricted-neighbourhood of p. (b) Core and noise points. (c) Multi-cluster membership. The three red points in the sub-figure (b) are *core* points whereas the ones coloured in green are classified as *noise*. In the sub-figure (c), the point coloured in blue is a member of two clusters: the red and the black clusters of points.

7.5.2 Detection of Social Groups

SocialDBC uses the concept of ϵ-restricted-neighbourhood and the thresholds: δ, ϵ and $minPts$ to classify the points given in D into *core* and *noise* points. Consequently, it links those core points that are *socially connected* into social groups. Figure 7.2 illustrates the concepts of ϵ-restricted-neighbourhood, the two classes of points: core and noise, as well as points for multi-cluster membership.

Algorithm 7.1 (see § 7.5.2) gives the pseudo code for SocialDBC, which starts by declaring an empty set of core points (line 2). It then performs three tasks for each point given in D: it computes the neighbourhood $N_\epsilon(p)$, if p satisfies the requirement for core points, it adds p to the set of core points, and then it declares that p is assigned to none of the social groups by setting the set of ids, belonging to p, as being empty (lines 3-9).

In the next step, for each core point with no cluster assignment, SocialDBC finds the set of socially connected points for the given core point (line 12). If the detected set size is greater than or equal to the threshold *delta*, the set is identified as a social group and as a

```
1: SocialDBC(D, ϵ, δ, minPts = 2)
2: Core ← ϕ
3: for each p ∈ D do
4:     Compute N_ϵ(p)
5:     if RN_ϵ(p) ≥ minPts then
6:         Core ← Core ∪ {p}
7:     end if
8:     id_p ← ϕ
9: end for
10: k ← 0
11: for each p ∈ Core do
12:     G ← FindSocialGroup(p, ϵ, δ)
13:     if |G| ≥ δ then
14:         k ← k + 1
15:         for each p ∈ G do
16:             id_p ← id_p ∪ {k}
17:         end for
18:     end if
19: end for
20: Groups ← {G_i | G_i = {p | p ∈ D, i ∈ id_p}}
21: Noise ← {p ∈ D | id_p = ϕ}
22: return Groups, Noise
23:
24: FindSocialGroup(p, ϵ)
25: ψ ← p
26: for each q ∈ N_ϵ(p) do
27:     set ConnectedPt ← True
28:     for each r ∈ ψ do
29:         if |θ_{q,r}| > ϵ then
30:             ConnectedPt ← False
31:             break
32:         end if
33:     end for
34:     if ConnectedPt = True then
35:         ψ ← ψ ∪ {q}
36:     end if
37: end for
38: return ψ
```

Algorithm 7.1: Social Density-based Clustering (SocialDBC)

result the set of ids associated with each point in the social group is amended to indicate that the point is a member of the newly detected social group.

A point may be connected to multiple social groups. Such a point is added to all of those social groups that the point is connected to. Any point that has not been assigned to a social group is considered to be *noise*.

7.5.3 SocialDBC vs DBSCAN

A major distinction between the proposed SocialDBC algorithm and the many DBSCAN versions that exist in the literature is that the former discovers only convex clusters [35]

of points. A fundamental concept of the social grouping discussed in this section is that detected social groups must not include *a-friend-of-a-friend* relationships, which DBSCAN inherently allows through the creation of elongated non-convex clusters [35].

Another subtle difference between the two methods manifests in how multi-cluster participation is perceived. Overlapping of clusters conforms with how social grouping is defined in this research, where an individual can be a member of multiple social groups irrespective of the type of social interaction. While the social groups discovered by SocialDBC are not exclusive, i.e. SocialDBC permits the participation of points in multiple clusters, DBSCAN and its two variations, i.e. *Social-DBSCAN* and *Temporally-Restricted-Social-DBSCAN* (see § 6.4 and § 6.5), produce exclusive clusters where overlapping is not permitted.

One important feature of the SocialDBC method is the two level computation of the distance between two points. In addition to using Jaccard distance [19] to find the neighbourhood of a given point, we apply a minimum number of visits threshold to filter out those neighbouring points that do not belong to the social group. For example, to detect the group of students that attend the same class, we first find all the students that are part of the neighbourhood of an observed student. To do this we compute the Jaccard distance between the set of locations that the observed student visited and the set of visited locations of each of the students recorded in the database [20]. From the obtained neighbourhood we further filter the group of students that made joint visits greater than or equal to a minimum threshold of joint visits. This group of students that meets the joint visit criterion is considered to be a social group.

The key limitation, which both methods share, is the sensitivity of the result of clustering to the value of ϵ, especially when the underlying clustering that we seek to discover has a wide range of density values.

The two methods have similar complexity due to the computation of the neighbourhood for each point in the data set. Thus, the worst-case complexity for SocialDBC is $O(n^2)$ [110].

7.6 Modelling Social Presence

7.6.1 The Social Presence Model (SPM)

We propose the SPM model, which classifies locations into *formal* and *informal* locations on the basis of the visiting patterns detected at those locations. We have learnt so far how social groups can be detected using spacial and temporal information extracted from Wi-Fi activity traces and we would like to formulate a model that exploits these visiting patterns to predict the type of location where people socialise.

Based on our definition of *formal social presence* in § 7.4.1, the visits made to an observed location by the same social group represent a set of uniformly distributed points in the visit space. Consequently, for each social group we can test for a discrete uniform distribution applied to the group's set of visits, recorded at the observed location. To illustrate the idea, we proceed as follows.

Given a location l, for each detected social group, we compute the length of the time period between each visit and the next. The data set made of these period lengths can be regarded as a sample s, which we hypothesise to be uniformly distributed. Formally, for each social group that visited the location l we find the set of visits $v_1, \ldots, v_n$, arranged in chronological order. We compute the number of days between each two consecutive visits to create the set s. We denote the set comprising all the sets of in-between visits gaps for the current location as S, thus $|S|$ denotes the number of social groups that visited the observed location. Assigning l to the class of *formal locations* can be estimated by counting how many sets $s \in S$ are approximately uniformly distributed. Therefore, the probability of the observed location l being classified as a *formal location* can be computed as the proportion of the number of uniformly distributed sets $s \in S$, compared to the number of social groups that visited the observed location, namely

$$\Pr\big(Y = formal\big) = \frac{\sum_{s \in S} I(s \sim U(a,b))}{|S|}, \tag{7.4}$$

where I is an indicator function that has value 1 only when its argument is true, and 0 otherwise; a and b are the minimum and maximum number of days between two consecutive visits.

Since we only have two types of location: *formal* and *informal*, classifying a location as *formal* corresponds to predicting that its type is *formal* if $\Pr\big(Y = formal\big) > \eta$, and

informal otherwise.

NB The minimum probability threshold η is a user-specified value from the interval [1,0) - the value 0.5 has worked very well in all experimentations.

To verify the uniformity of $s \in S$, we use the following hypotheses:

H_0: The periods lengths in s are uniformly distributed.
H_1: The periods lengths in s are not uniformly distributed.

In order to test these hypotheses, we compute the chi-square goodness of fit statistic as shown below [50, 61].

$$T = \frac{\sum_{i=1}^{d}(O_i - E_i)^2}{E_i} \approx \chi^2_{d-1} \tag{7.5}$$

where O_i is the observed count of the period length i, E_i denotes the expected count, $E_i = \frac{1}{|s|}\sum_{j=1}^{d} O_j$, and d is the number of count values O_i based on the observed s.

7.6.2 Baseline Model

We use a multiple logistic regression model as a baseline model for comparison. The model infers the type of an observed location based on a set of features, which describe each social group that attended the location: the size of the group, number of visits made by the group, minimum and maximum number of days between two consecutive visits. It is a global model in the sense that the model is fitted using information from all formal and informal locations in our data. We estimate the probability of whether an observed location can be classified as *formal* or *informal* using the following equation.

$$\Pr\big(Y = formal\big) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4}}, \tag{7.6}$$

The maximum likelihood method [43] is used to estimate the parameters $\beta_0,\ \beta_1, \ldots,\ \beta_4$. $x_1,\ x_2,\ x_3\, x_4$ denote the size of the group, number of visits made by the group, and the minimum and maximum number of days between two consecutive visits, respectively.

An observed location is classified as *formal* if $\Pr\big(Y = formal\big) > 0.5$, and *informal* otherwise, namely (≤ 0.5).

7.7 Evaluation

7.7.1 Data Set

The evaluation of the proposed clustering method, Social-DBSCAN, is based on recent WLAN traces collected at Birkbeck - the case-study university in this research work. A detailed description of the evaluation data has been provided in § 5.6.1.

7.7.2 Experiments

We evaluate the proposed SPM[1] and baseline models on the eduroam data set, which we describe in section 5.6.1 of the previous chapter. We are particularly interested in the predictive performance of the models, i.e., given the information about the visits made by different social groups, our goal is to accurately predict the type of each location visited.

In order to detect students social groups, which we subsequently used in the evaluation of our proposed prediction models, the raw eduroam data was processed to create m-dimensional points. Each point denotes the *visits* made, by one of the users, to different locations across the university campus.

7.7.2.1 Evaluation Metrics

In order to measure the performance, we consider the mean prediction accuracy as an evaluation metric, i.e. an accuracy of 0.1 means that only 10% of the time the proposed model, namely the SPM model and the baseline model, successfully predicts the correct type of the observed location. Also for each model, we provide a table of confusion (a confusion matrix) to report the number of false positives, false negatives, true positives, and true negatives. The significance levels of 0.01 and 0.05 were used for performing the statistical hypothesis testing.

7.7.2.2 Experimental Setup

Our experiments were based only on detected social groups which visited the set of locations given in Table 7.1. These social groups were detected using our proposed clustering method SocialDBC. Each social group had a least two visits to the same observed location. To evaluate the accuracy of the SPM model, we used all the data without division into training and testing data sets. However, for the evaluation of the baseline model (see

[1]The Social Presence Model

§ 7.6.2) we divided the data into training and testing sets, where from each location we used 80% of data for training and the remaining 20% for testing.

7.7.3 Results

7.7.4 Uniformity of Social Presence Across *Formal* Locations

Our initial intuition is that the locations associated with formal activities are visited in a regular manner by social groups with uniform periods between visits. In contrast, those locations that are linked to informal activities have irregular patterns of visits. To verify this intuition, we evaluated the proposed models, described in § 7.6, on the visit data of each of the locations given in Table 7.1.

		Actual Location Type	
	Predicted Location Type	Formal	Informal
SPM^1	Formal	8	1
	Informal	0	3
SPM^2	Formal	8	1
	Informal	0	3
Baseline	Formal	8	4
	Informal	0	0

Table 7.3: Table of Confusion. SPM^1 and SPM^2 represent the SPM model using the significance levels of 0.01 and 0.05, respectively.

Table. 7.3 reports the number of false positives, false negatives, true positives, and true negatives from the evaluation of these two models. The reported results show how the SPM model (i.e. the SPM^1 and SPM^2 which represent the SPM model using the significance levels of 0.01 and 0.05, respectively) offers a significantly improved performance over the baseline model, which they outperform by a factor of 1.38 in terms of accuracy: 0.92 for the SPM^1, and the SPM^2 models, and 0.67 for the baseline model. Both SPM models,i.e. SPM^1 and SPM^2, correctly classified three informal locations out of four whereas the baseline model failed to correctly classify any of them. Interestingly, the two SPM models and baseline model correctly classified all eight formal locations. Figure 7.3, plots the classification of locations into formal and informal locations based on the predictions made by the two SPM models and the baseline model, as shown in the sub-figures (a), (b) and (c) respectively.

7.7.5 The Geographical Spread of Social Meetings Across Campus

We studied the number of locations visited by social groups across campus and we discovered that around 83% of the detected groups visited only one location to socialise. Figure 7.4, shows the distribution of the number of locations visited by social groups across locations where informal activities occur. We also examined the number of locations visited by social groups that attended the Coffee-shop and the Bar at Malet Street and for each group we counted the number of visited locations from other sites of the campus, i.e. places located off-site Malet Street. As shown in Figure 7.5, 91% and 99% of social groups that visited the informal locations at Malet Street and Gordon Square, restricted their visits to nearby locations, i.e. locations within the same site, as opposed to locations that are further afield. One interpretation of such result is that many social groups visit *informal locations* at lunchtime and in coffee breaks during lectures and other learning sessions. These breaks usually last for short periods, and consequently do not provide enough time for groups to socialise off-site far from their prime location of work or study.

7.7.6 Visiting Behaviour Across Locations

Intuitively, formal locations, where activities such as learning classes and lab sessions take place, are usually attended by groups as opposed to individual users. To find out whether users visit a given location as a group or individually, we calculate the *social weight*, which compares the number of shared visits made by the social group to the total number of visits made by the individual user, including the visits they made with their social group:

$$SocialWeight = \frac{Number\ of\ group\ visits}{Number\ of\ individual\ user\ visits} \tag{7.7}$$

In ideal settings, a *social weight* value that is equal/close to 1 demonstrates the superiority of group visits over the individual user visits. In contrast, a significantly smaller value is a clear indication that the user prefers to visit the observed location as an individual as opposed to visiting it with a group. Figure 7.6 illustrates such scenarios where the skewness of the distribution indicates the dominance of one type of visiting behaviour over the other.

As shown in Figures 7.7 and 7.8, the *social weight* value varies from one observed

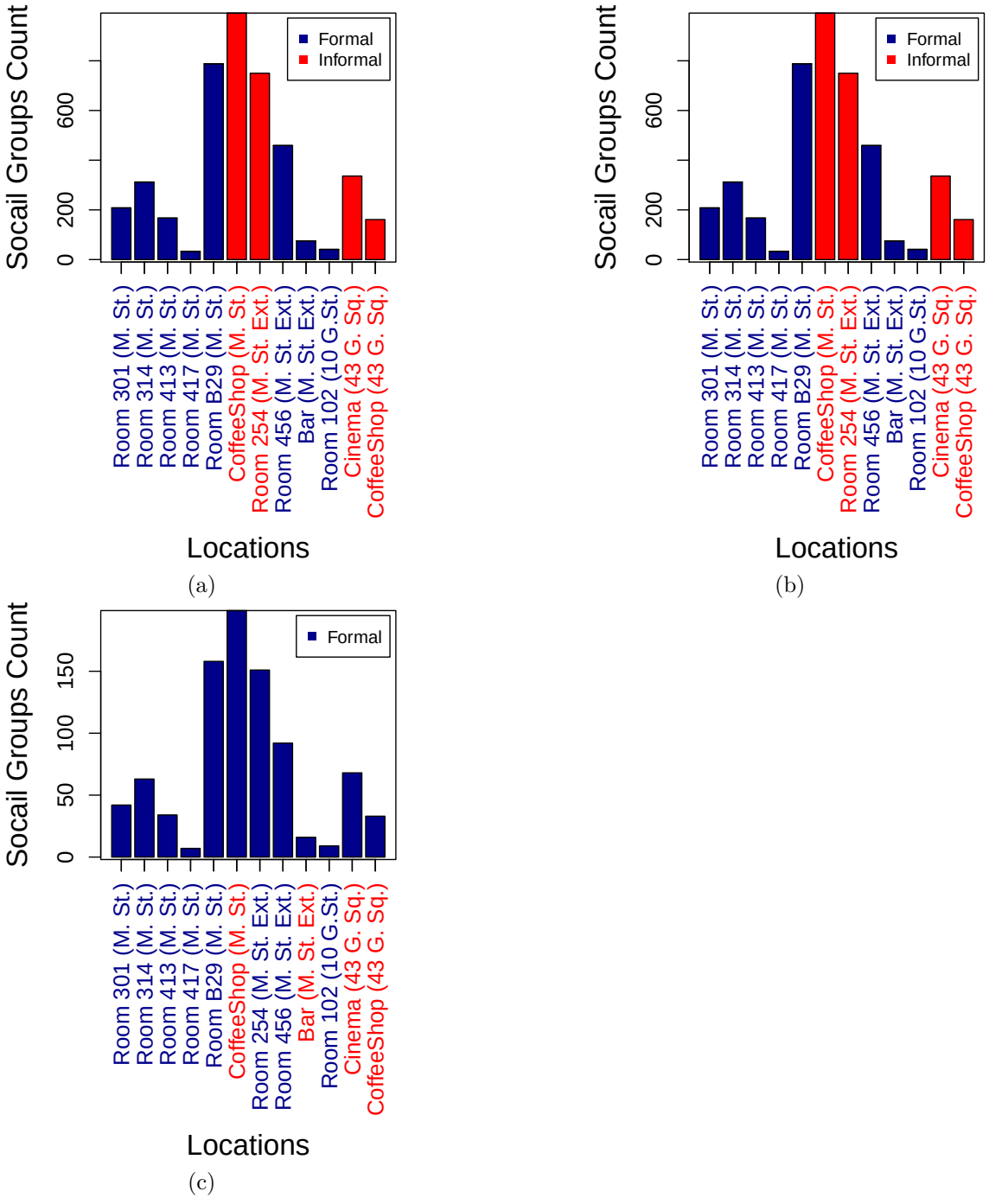


Figure 7.3: Classification of locations into formal and informal locations based on the predictions made by (a) the SPM model using a significance level of 0.01, (b) the SPM model using a significance level of 0.05 and (c) the baseline model. The colours of plotted location names reflect the two types of location given in Table 7.1.

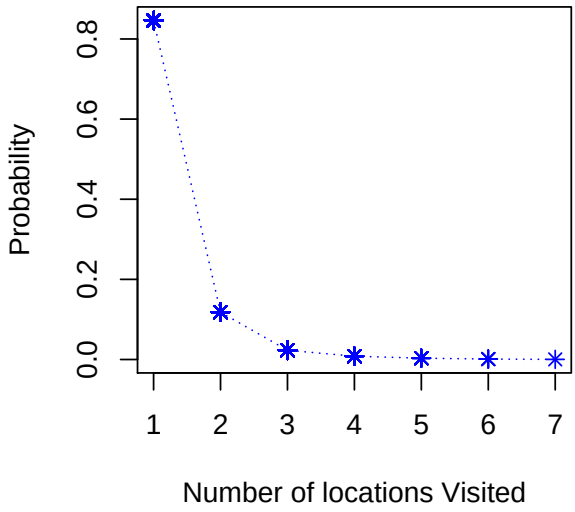
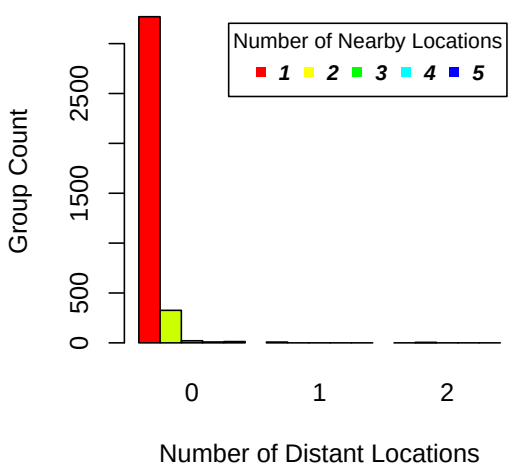
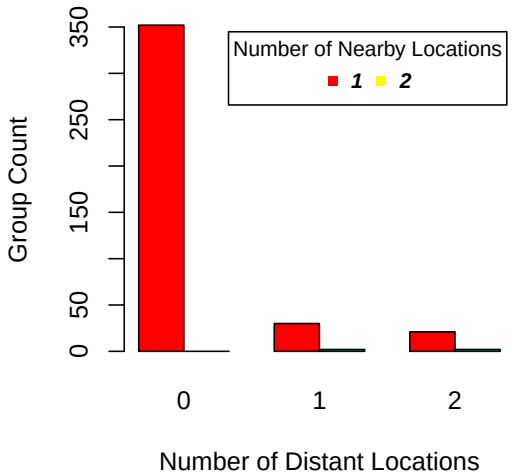


Figure 7.4: Distribution of number of locations visited by social groups detected across all locations.



(a) Malet Steet



(b) Gordon Square

Figure 7.5: Number of distant locations visited by social groups that visited Malet Street and Gordon Square informal locations. In this experiment, a distant location is any Birkbeck location excluding the ones situated at Malet Street and Gordon Square.

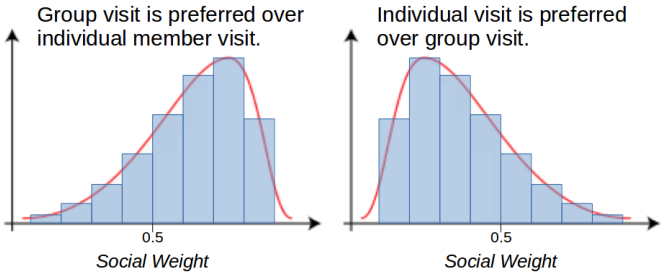


Figure 7.6: Types of visiting behaviour as seen through the distribution of ratio of number of group visits compared to the number of individual member visits.

location to another but generally those locations which are linked with formal activities seem to be favoured by social groups as opposed to individual users. With exception of the distribution for Room B11 at 43 Gordon Square site, it is clearly evident from the negative skewness of the peaked distributions shown in Figure 7.7 that more visits were made, to these locations, by social groups as opposed to individual users. Although the distribution for Room B11 has a positive skewness but the social weight values shown range between 0.6 and 1.0, which clearly indicates that the location was visited by groups of users more than it was visited by individual users.

Similar to formal activity locations, most of the observed locations associated with informal activities seem to have the group behaviour of visit as the favoured mode of visit. As shown in Figure 7.8, the negatively skewed and highly peaked distributions for locations such as the Coffee Shops suggest that they are preferred locations for social groups. Despite the positive skewness of distribution for the Bar at Malet Street Extension, the social weight values shown are greater than 0.5, which strongly indicates that the location was visited by groups of users more than it was visited by individual users. The Cinema at 43 Gordon Square seems to have a large proportion of its visits made by individual users but it nonetheless remains a favoured destination for social groups.

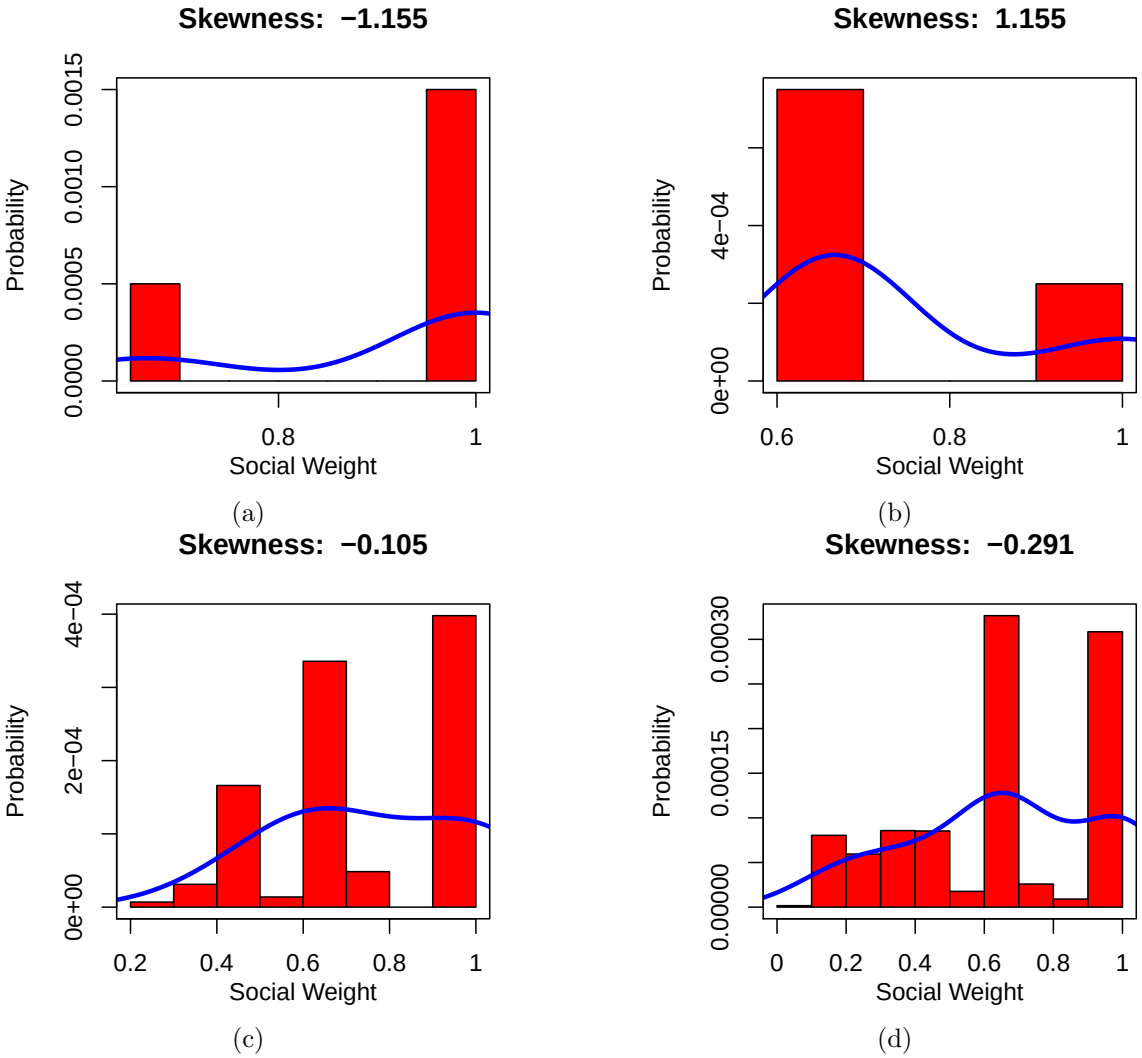


Figure 7.7: Distributions of *social weight* for formal activity locations. Shown from left to right and from top to bottom are the distributions for: (a) Room 102 at 10 Gower Street, (b) Room B11 at 43 Gordon Square, (c) Room 314 at Malet Street and (d) Room 254 at Malet Street Extension.

7.8 Discussion

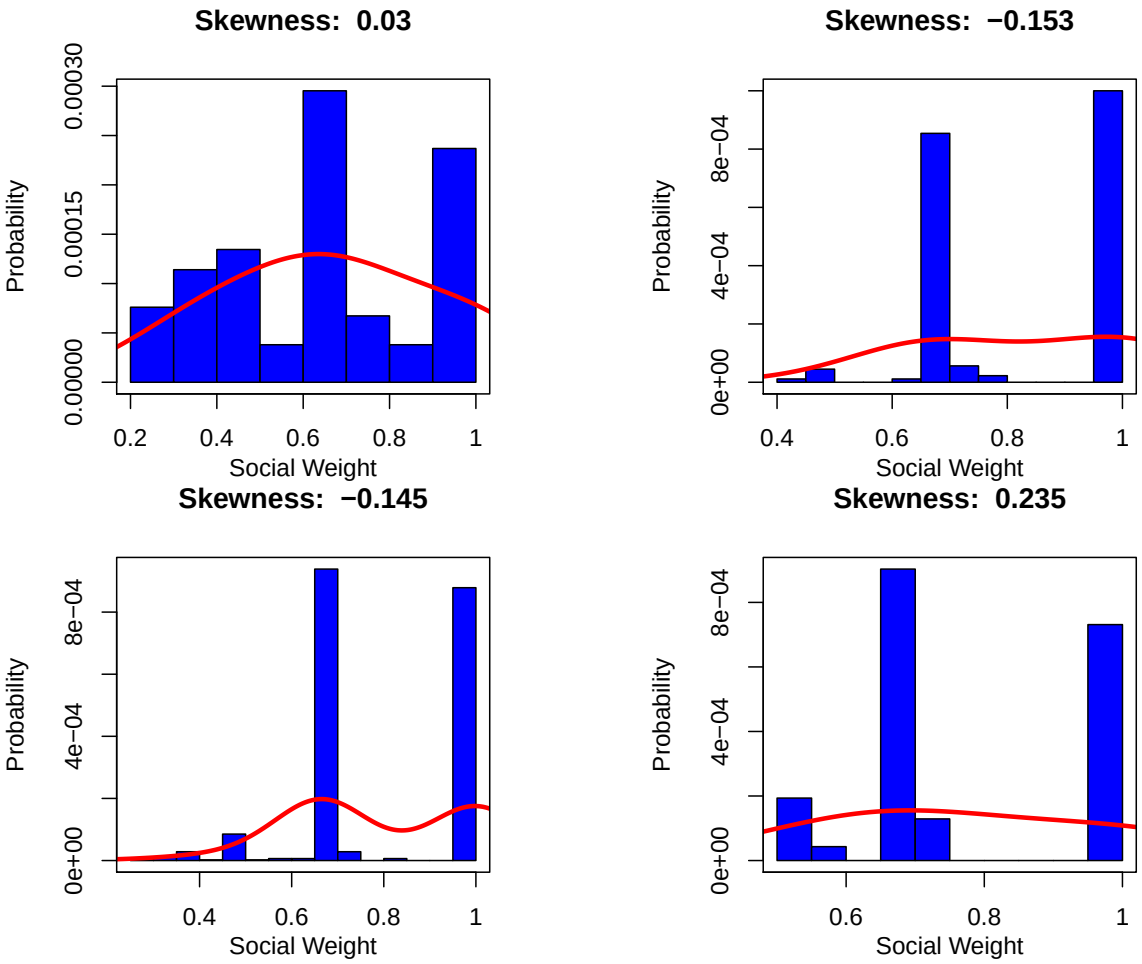
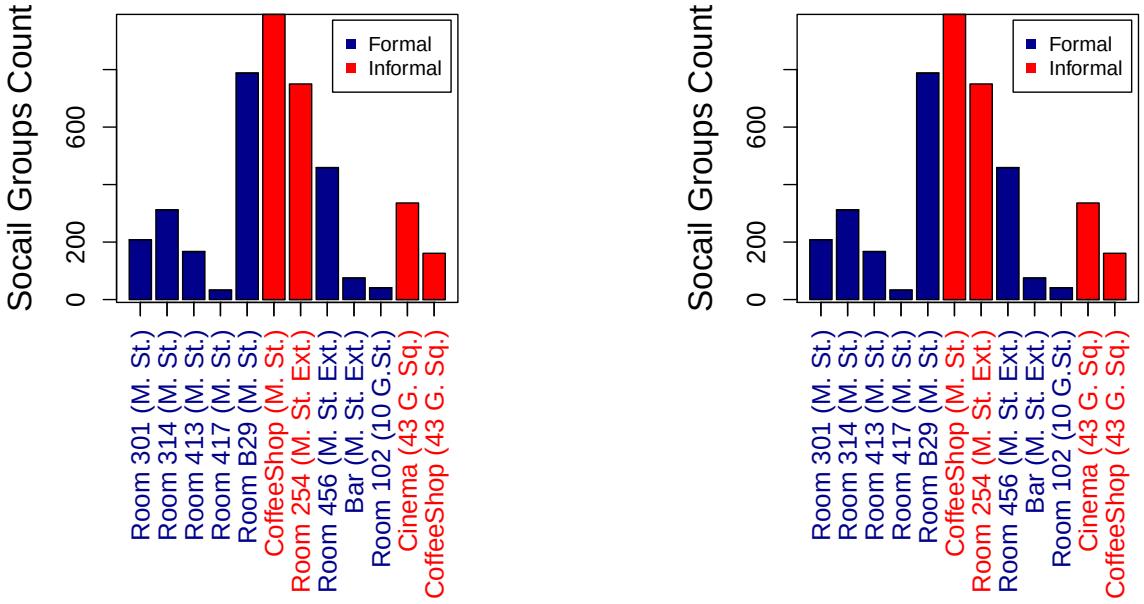


Figure 7.8: Distributions of *social weight* for informal activity locations. Shown from left to right and from top to bottom are the distributions for: the Cinema, the CoffeeShop at 43 Gordon Square, the Coffee Shop and the Bar at Malet Street.

7.8.1 Chi-square Test

An important aspect of the SPM performance to report herein is when all $E_i \geq 1$, and at least 80% of them ≥ 5. This is a sound application of the Chi-square test which we would like to perform in this analysis. As shown in Figure 7.9, the reported results under such a condition, namely all $E_i \geq 1$, and at least 80% of them ≥ 5, still show the improved performance of the SPM model (i.e. the SPM^3 and SPM^4 which represent the SPM model using the significance levels of 0.01 and 0.05, respectively) over the baseline model, where the two SPM models maintained the same performance in terms of accuracy: 0.92 for the

SPM^3, and the SPM^4 models.



(a) (b)

Figure 7.9: Classification of locations into formal and informal locations based on the predictions made by (a) the SPM model using a significance level of 0.01, and (b) the SPM model using a significance level of 0.05. In this experiment all $E_i \geq 1$, and at least 80% of them ≥ 5. The colours of plotted location names reflect the two types of location given in Table 7.1.

7.8.2 Lack of Proximity Data

Similar to the discussion provided in Section 6.7.6 in the previous chapter, the evaluation data we utilised in this research does not comprise proximity information between two (or more) co-located individuals visiting an observed location - for example the distance between two individuals visiting the Coffee-shop but sitting at separate tables. Although the proposed clustering method does not employ such granular proximity information in deciding the membership of a social group, it is most likely that by utilising such information, achieving a higher accuracy in discovering social groups is feasible.

7.9 Summary

The key contributions of this chapter can be summarised as follows:

1. We demonstrated how by clustering WLAN activity traces we can detect social groups of mobile users within an academic environment. Moreover, we showed how by being able to detect social groups at target locations, we provide an invaluable opportunity to understand the presence and movement of people within such an environment.

2. We developed a clustering method, which we call SocialDBC, that leverages on the type of activity performed at an observed location in order to detect visiting social groups. We discovered that people generally socialise at a very small set of nearby locations within campus - within the same building or site. Generally, people visited a distant location, i.e. another Birkbeck site, when they were in the company of their social group.

3. Given the categorisation of occupied spaces into two main types: formal and informal locations, our proposed model of human social presence (SPM) can infer the type of any observed location based on the visiting behaviours exhibited at that location. This seemingly simple model reliably predicts the correct visited location type and offers significantly improved performance over the nontrivial baseline model which failed to make a correct prediction when the location type is informal.

Chapter 8

Conclusions

In this thesis, we studied three main aspects of the human presence and movement behaviour within specific environments: spatio-temporal movement (where and when do people move), user identification (how to uniquely identify people from their presence and movement historical records), and social grouping (how do people interact). We considered two environments: a learning environment represented by a university campus and a city environment represented by an average-size city in Europe. The two large data sets that we utilised in the evaluation of the models described herein capture the presence and movement behaviour in these two environments. Employing these two data sets, we investigated the three aspects of the human presence and movement behaviour which are summarised hereafter.

8.1 Summary of the Thesis

The contributions of this thesis are divided into two parts: the *first part*, investigates the spatio-temporal movement where we predict the future locations of visit based on when and where users had been in the past. We also investigate the possibility of using recorded movements for *user identification*. The stochastic models proposed for *movement prediction* and user identification were evaluated on the data set obtained from Nokia (see Subsection 1.6.1.1).

Chapter 3. Considering the next location prediction problem, the one-model-per-user and the *collective* model approaches were investigated. The two approaches were found to have very comparable prediction performances, particularly when previously seen behaviours are available to make inferences from. Furthermore, the effect of the length of

the user record of the most recent temporal locality used in predicting the next location of visit was examined. It was shown that only a short record of mobility history is required in order to make relatively accurate predictions about future locations of visit. Moreover, the effect of the length of this record and the relative loss of accuracy when reduced data samples are used was investigated. It was clear from the results of the experiments we carried out that as the length of the historical record increases the models' prediction accuracy improves.

The proposed *collective* approach has the potential to overcome the one-model-per-user's weaknesses such as the inability to deal with novel behaviours.

The performance of the proposed prediction models, i.e. the single user models and the collective multi-user models, were evaluated by using MAE and RMSE error metrics [13]. It was shown that these two metrics can be utilised in computing the suggested (the top-k) locations which are most likely to include the observed user's correct next location of visit. The merits of HM Score in assessing the accuracy of the proposed models were examined and employing HM score and the two mean error metrics, i.e. the *Mean Absolute Error* and the *Root Mean Square Error*, seem to provide a broader view of the prediction accuracy as opposed to applying a single metric.

Chapter 4. The *mobility fingerprint* [32], which is a profile constructed from the user's historical mobility traces was proposed. An algorithm for building such a profile, which we evaluated by collecting a sample of fingerprints from the publicly available Nokia Mobile Data Challenge data set (see § 1.6.1.1) was introduced. Furthermore, it was shown that users have unique mobility fingerprints, i.e. they can be distinguished from one another based on their mobility fingerprints. Moreover, an observed mobility trail can be associated with the fingerprint of the user to whom the trail belongs, i.e. a user can be identified by his/her movements. We showed that in order to successfully identify individual users on the basis of their recent mobility history, it is imperative that a rich historical record about the movement of those users is maintained. It was shown herein that the richer the fingerprint the more accurate the identification of the user from observed movements is. Also the idea of whether the proposed fingerprinting method [32] can be extended to create unique profiles for landmarks and whether such fingerprints can be used for location prediction was explored. To this end, it was shown that the proposed fingerprinting method can be used to create unique profiles for landmarks and successfully be employed

in the context of the *Next Location Prediction problem*.

In the *second part* of the thesis we considered the concept of social grouping (how do people interact). The clustering models proposed for the detection of social groups, and location classification, within an observed learning environment, were successfully applied to the Eduroam data (see § 1.6.1.2) obtained from Birkbeck, University of London.

Chapter 5. A comprehensive analysis about the human presence within a university campus was carried out where a thorough analysis about the four types of patterns contained in the data: the social, the spatial, the temporal and the semantic patterns, was provided. For each of these types of pattern: the social, the spatial, the temporal and the semantic, we defined a list of metrics in order to interpret the observed behaviour captured in the data, and thus giving an insight into how people presence shapes the dynamic structure of such an environment.

Chapter 6. Two social density-based clustering methods that utilise WLAN traces in order to detect granular social groups of mobile users within a university campus were proposed. These clustering methods rely on the underpinning semantic context for parameterisation, i.e. utilise information from the semantic context to determine the values of the parameters of the proposed clustering algorithms. The actual level of attendance of learning activities was estimated by linking the discovered social group that regularly visits an observed location and the learning activity that takes place within the same context.

Chapter 7. A density-based clustering method [33] that discovers social groups by utilising activity traces of mobile users was introduced. The proposed algorithm was successfully applied in detecting the social groups on the basis of the activities taking place at observed locations within a university campus. Furthermore, a framework for inferring the type of an observed location, by using the patterns of visit extracted from Wi-Fi activity traces was proposed, and implemented [33].

8.2 Summary of Contributions

Overall, the thesis makes the following contributions:

- A novel family of predictive models that allows for inference of locations though a collaborative mechanism which does not require the profiling of individual users.

These prediction models utilise suffix trees as their core underlying data structure, where predictions about a specific individual are computed over an aggregate model incorporating the collective record of observed behaviours of multiple users.

- A novel approach for user identification which we call *mobility fingerprint*. This user identification method is a profile constructed from the users historical mobility traces. The proposed method for constructing such a profile is a principled and scalable implementation of a variable length Markov model based on n-grams.
- Novel density-based clustering methods that discover social groups by analysing activity traces of mobile users as they move around, from one location to another, within an observed environment.
- A novel framework for inferring the type of an observed location within a university environment, by using the patterns of visit extracted from Wi-Fi activity traces.

8.3 Constraints and Limitations

The key constraints and limitations of this thesis are briefly highlighted hereafter.

8.3.1 Lack of Proximity Information

Measuring the proximity between two co-located individuals visiting an observed location, is essential to accurately predict whether or not those two individuals are there to socialise. For example, if two students are visiting the Coffee-shop and we do not have information about the distance between their two seats, it would be hard to decide whether they are sitting at the same table or at separate tables. Unfortunately, the Wi-Fi data set we utilised for the evaluation of the social groups detection methods proposed in this thesis does not contain such proximity information; consequently it is hard to accurately determine if two people are engaged in a one-to-one or any other form of social interaction that requires participating individuals to be within close distance from one another.

A novel system which enables a single Wi-Fi *access point* to localise devices within a proximity of tens of centimetres range was proposed in [102]. With the aid of such a system it is possible to obtain a rich data set that includes information about the proximity between users visiting an observed location. Since we are constrained by the quality of the data set utilised herein, the proposed social group detection models in this thesis do

not take the proximity between users into consideration when detecting social groups at an observed location.

8.3.2 Lack of Ground Truths

Lack of rich record of ground truths of courses taught across the university including attendance records of individual learning sessions and social meetings outside the classroom is a limitation in this work. Evaluating our proposed methods using a data set that include such ground truths provides an invaluable opportunity to carry out a robust assessment of our proposed methods in this thesis. Such an assessment will provide new insight about the accuracy of the social groups detection models proposed herein.

8.4 Future Research Directions

Our future research is threefold. Although this future research is indicated in terms of a learning environment, the ideas can be applied to other environments such as an organisation office complex.

8.4.1 The Influence of Presence and Mobility Behaviour on Academic Performance

We are increasingly adopting a life style in which intensive interactions and communication using different devices over the Internet is part of our daily routine. The ability to link data of different types from across the information ecosystem including social media, location-based networks, telephone service providers, smart cities and wearable devices, provides an opportunity to capture and to study the human presence and mobility behaviour in much more depth as never before. With such data arrangement in place, we can study the influence of various factors of the human presence and mobility behaviour within a given environment. For example, one of our research interests is to study the influence of the different aspects of the human presence and movement behaviour, namely the spatial, the social, and the temporal aspects, on the academic performance of students within a given learning environment.

8.4.2 Estimating Spatial Occupancy

There has been a growing interest in exploiting existing technologies, such as Wi-Fi, in order to track the human presence and movement behaviour at an observed university

campus. Utilising an existing Wi-Fi network in tracking attendance has a direct benefit in cutting costs, particularly when other specialised tracking technologies are considered. For example, using video camera based techniques for counting occupants [60] where usually there are significant installation and running costs involved in using such specialised technologies. Developing a model that utilises an existing Wi-Fi network, for estimating the actual space usage linked to *formal* and to *informal* activities that take place at a target university environment [33], is a key topic of our current research interests.

8.4.3 Reducing the Size of a Mobility Fingerprint

We would like also to undertake further investigation in order to discover the true implications on the user identifiability when the mobility fingerprint is compressed. We would like to carry out this investigation using, at least, one additional large data set. We would also like to investigate whether the location fingerprints can be used for location recommendation. A key part of our future research will be dedicated to the idea of enriching the fingerprints with additional information so as to gain better understanding of the user's interests and personal tastes.

Bibliography

[1] Anisa Allahdadi, Ricardo Morla, Ana Aguiar, and Jaime S Cardoso. Predicting short 802.11 sessions from radius usage data. In *38th Annual IEEE Conference on Local Computer Networks-Workshops*, pages 1–8. IEEE, 2013.

[2] Raul Amici, Marco Bonola, Lorenzo Bracciale, Antonello Rabuffi, Pierpaolo Loreti, and Giuseppe Bianchi. Performance assessment of an epidemic protocol in vanet using real traces. *Procedia Computer Science*, 40:92–99, 2014.

[3] Nguyen Thanh An and Tu Minh Phuong. A gaussian mixture model for mobile location prediction. In *2007 IEEE International Conference on Research, Innovation and Vision for the Future*, pages 152–157. IEEE, 2007.

[4] Daniel Ashbrook and Thad Starner. Learning significant locations and predicting user movement with gps. In *Proceedings. Sixth International Symposium on Wearable Computers,*, pages 101–108. IEEE, 2002.

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

[6] James P Bagrow, Dashun Wang, and Albert-Laszlo Barabasi. Collective response of human populations to large-scale emergencies. *PloS One*, 6(3):e17680, 2011.

[7] Anand Balachandran, Geoffrey M Voelker, Paramvir Bahl, and P Venkat Rangan. Characterizing user behavior and network performance in a public wireless lan. In *ACM SIGMETRICS Performance Evaluation Review*, volume 30, pages 195–205. ACM, 2002.

[8] Bharathan Balaji, Jian Xu, Anthony Nwokafor, Rajesh Gupta, and Yuvraj Agarwal. Sentinel: Occupancy based hvac actuation using existing wifi infrastructure within

commercial buildings. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, page 17. ACM, 2013.

[9] Jie Bao, Yu Zheng, and Mohamed F Mokbel. Location-based and preference-aware recommendation using sparse geo-social networking data. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, pages 199–208. ACM, 2012.

[10] Gill Bejerano and Golan Yona. Variations on probabilistic suffix trees: Statistical modeling and prediction of protein families. *Bioinformatics*, 17(1):23–43, January 2001.

[11] Brigitte Bigi. *Using Kullback-Leibler Distance for Text Categorization.* Springer, 2003.

[12] Jose Borges and Mark Levene. Evaluating variable-length markov chain models for analysis of user web navigation sessions. *Knowledge and Data Engineering, IEEE Transactions On*, 19(4):441452, 2007.

[13] Jose Borges and Mark Levene. A comparison of scoring metrics for predicting the next navigation step. *November*, 3(2010):15, 2010.

[14] José Borges and Mark Levene. A comparison of scoring metrics for predicting the next navigation step with markov model-based systems. *International Journal of Information Technology & Decision Making*, 9(04):547–573, 2010.

[15] Kendrick Boyd, Kevin H Eng, and C David Page. Area under the precision-recall curve: Point estimates and confidence intervals. In *Machine Learning and Knowledge Discovery in Databases*, pages 451–466. Springer, 2013.

[16] Robert H Bruininks, Brianne Keeney, and Jim Thorp. Transforming americas universities to compete in the new normal. *Innovative Higher Education*, 35(2):113–125, 2010.

[17] Xiao Cai, Feiping Nie, and Heng Huang. Multi-view k-means clustering on big data. In *Twenty-Third International Joint Conference on Artificial Intelligence*, pages 2598–2604. AAAI, 2013.

[18] Xin Cao, Gao Cong, and Christian S Jensen. Mining significant semantic locations from gps data. *Proceedings of the VLDB Endowment*, 3(1-2):1009–1020, 2010.

[19] Sung-Hyuk Cha. Comprehensive survey on distance/similarity measures between probability density functions. *City*, 1(2):1, 2007.

[20] Moses S Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the Thiry-fourth Annual ACM Symposium on Theory of Computing*, pages 380–388. ACM, 2002.

[21] Alan H Cheetham and Joseph E Hazel. Binary (presence-absence) similarity coefficients. *Journal of Paleontology*, pages 1130–1136, 1969.

[22] Kai San Choi, Edmund Y Lam, and Kenneth KY Wong. Source camera identification using footprints from lens aberration. In *Digital Photography II*, volume 6069, pages 172–179, 2006.

[23] Yohan Chon, Hyojeong Shin, Elmurod Talipov, and Hojung Cha. Evaluating mobility models for temporal prediction with high-granularity mobility data. In *2012 IEEE International Conference on Pervasive Computing and Communications*, pages 206–212. IEEE, 2012.

[24] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[25] Justin Cranshaw, Eran Toch, Jason Hong, Aniket Kittur, and Norman Sadeh. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, pages 119–128. ACM, 2010.

[26] Nuno Cruz, Hugo Miranda, and Pedro Ribeiro. The evolution of user mobility on the eduroam network. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference On*, pages 249–253. IEEE, 2014.

[27] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, 3:1376, 2013.

[28] Nathan Eagle and Alex Sandy Pentland. Reality mining: Sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006.

[29] Peter Eckersley. How unique is your web browser? In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 1–18. Springer, 2010.

[30] Muawya H Sarnoub Eldaw, Mark Levene, and George Roussos. Presence analytics: Detecting classroom-based social patterns using wlan traces. In *2017 Intelligent Systems Conference (IntelliSys)*, pages 346–353. IEEE, 2017.

[31] Muawya Habib Sarnoub Eldaw, Mark Levene, and George Roussos. Collective suffix tree-based models for location prediction. In *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication*, pages 441–450. ACM, 2013.

[32] Muawya Habib Sarnoub Eldaw, Mark Levene, and George Roussos. Poster: Constructing a unique profile for mobile user identification in location recommendation systems. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*, pages 479–479. ACM, 2015.

[33] Muawya Habib Sarnoub Eldaw, Mark Levene, and George Roussos. Presence analytics: Making sense of human social presence within a learning environment. In *2018 IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT)*, pages 174–183. IEEE, 2018.

[34] Dominik Maria Endres and Johannes E. Schindelin. A new metric for probability distributions. *Information Theory, IEEE Transactions On*, 49(7):1858–1860, 2003.

[35] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, and Others. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231. AAAI, 1996.

[36] Martin Ester, Hans-Peter Kriegel, Jrg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, volume 1996, page 226231, 1996.

[37] Vincent Etter, Mohamed Kafsi, and Ehsan Kazemi. Been there, done that: What your mobility traces reveal about your behavior. Technical report, 2012.

[38] Vincent Etter, Mohamed Kafsi, and Ehsan Kazemi. Been there, done that: What your mobility traces reveal about your behavior. In *Nokia Mobile Data Challenge 2012 Workshop. P. Dedicated Task*, volume 2, 2012.

[39] Adil Fahad, Najlaa Alshatri, Zahir Tari, Abdullah Alamri, Ibrahim Khalil, Albert Y Zomaya, Sebti Foufou, and Abdelaziz Bouras. A survey of clustering algorithms for

big data: Taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing*, 2(3):267–279, 2014.

[40] Trevor Fenner, Mark Levene, and George Loizou. A model for collaboration networks giving rise to a power-law distribution with an exponential cutoff. *Social Networks*, 29(1):70–80, 2007.

[41] Vernor C Finch. Geographical science and social philosophy. *Annals of the Association of American Geographers*, 29(1):1–28, 1939.

[42] Licia Florio and Klaas Wierenga. Eduroam, providing mobility for roaming users. In *Proceedings of the EUNIS 2005 Conference, Manchester*, 2005.

[43] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning*, volume 1. Springer Series in Statistics New York, 2001.

[44] Thomas George and Srujana Merugu. A scalable collaborative filtering framework based on co-clustering. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 4–pp. IEEE, 2005.

[45] Petko Ivanov Georgiev, Anastasios Noulas, and Cecilia Mascolo. The call of the crowd: Event participation in location-based social services. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[46] Alison L. Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

[47] Colin S Gillespie. Fitting heavy tailed distributions: The powerlaw package. *arXiv Preprint ArXiv:1407.3492*, 2014.

[48] Reginald G. Golledge. *Wayfinding Behavior: Cognitive Mapping and Other Spatial Processes*. JHU Press, 1999.

[49] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *nature*, 453(7196):779, 2008.

[50] Priscilla E Greenwood and Michael S Nikulin. *A Guide to Chi-squared Testing*, volume 280. John Wiley & Sons, 1996.

[51] Dan Gusfield. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge University Press, May 1997.

[52] Bruce E Hansen. Time series analysis james d. hamilton princeton university press, 1994. *Econometric Theory*, 11(3):625–630, 1995.

[53] Richard Hartshorne. B. is the limitation logically founded? *Annals of the Association of American Geographers*, 29(3):193–201, 1939.

[54] Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, and Samee Ullah Khan. The rise of big data on cloud computing: Review and open research issues. *Information Systems*, 47:98–115, 2015.

[55] Brian Hayes. Uniquely me! *American Scientist*, 102(2):106–109, 2014.

[56] Niall Hegarty. Where we are now–the presence and importance of international students to universities in the united states. *Journal of International Students*, 4(3):223–235, 2014.

[57] Tristan Henderson, David Kotz, and Ilya Abyzov. The changing usage of a mature campus-wide wireless network. *Computer Networks*, 52(14):2690–2712, 2008.

[58] Ordway Hilton. The complexities of identifying the modern typewriter. *Journal of Forensic Science*, 17(4):579–585, 1972.

[59] Hande Hong, Chengwen Luo, and Mun Choon Chan. Socialprobe: Understanding social interaction through passive wifi monitoring. In *Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pages 94–103. ACM, 2016.

[60] Ya-Li Hou and Grantham KH Pang. People counting and human detection in a challenging situation. *IEEE Transactions on Systems, Man, and Cybernetics-part A: Systems and Humans*, 41(1):24–33, 2010.

[61] Catherine Huber-Carol, Narayanaswamy Balakrishnan, M Nikulin, and M Mesbah. *Goodness-of-fit Tests and Model Validity*. Springer Science & Business Media, 2012.

[62] Jack D Ives and Bruno Messerli. Mountain hazards mapping in nepal introduction to an applied mountain research project. *Mountain Research and Development*, pages 223–230, 1981.

[63] P. Jacquet, W. Szpankowski, and I. Apostol. A universal predictor based on pattern matching. *Information Theory, IEEE Transactions On*, 48(6):14621472, 2002.

[64] Gudlaugur Jóhannesson, Gunnlaugur Björnsson, and Einar H Gudmundsson. Afterglow light curves and broken power laws: A statistical study. *The Astrophysical Journal Letters*, 640(1):L5, 2006.

[65] Minkyong Kim and David Kotz. Modeling users' mobility among wifi access points. In *Papers Presented at the 2005 Workshop on Wireless Traffic Measurements and Modeling*, pages 19–24. USENIX Association, 2005.

[66] David Kotz and Kobby Essien. Analysis of a campus-wide wireless network. *Wireless Networks*, 11(1-2):115–133, 2005.

[67] Hans-Peter Kriegel, Peer Kröger, Jörg Sander, and Arthur Zimek. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):231–240, 2011.

[68] John Krumm. A markov model for driver turn prediction. 2016.

[69] Frédéric Lassabe, Philippe Canalda, Pascal Chatonnay, François Spies, Numérica-Multimedia Developpement Center, and D Charlet. Predictive mobility models based on kth markov models. In *ICPS*, pages 303–306, 2006.

[70] J. K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T. M. T. Do, O. Dousse, J. Eberle, and M. Miettinen. The mobile data challenge: Big data for mobile computing research. In *Mobile Data Challenge By Nokia Workshop, in Conjunction with Int. Conf. on Pervasive Computing, Newcastle, UK*, 2012.

[71] Juha K Laurila, Daniel Gatica-Perez, Imad Aad, Olivier Bornet, Trinh-Minh-Tri Do, Olivier Dousse, Julien Eberle, Markus Miettinen, and Others. The mobile data challenge: Big data for mobile computing research. Technical report, 2012.

[72] Jong-Kwon Lee and Jennifer C Hou. Modeling steady-state and transient behaviors of user mobility: Formulation, analysis, and application. In *Proceedings of the 7th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pages 85–96. ACM, 2006.

[73] Jan Lukáš, Jessica Fridrich, and Miroslav Goljan. Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security*, 1(2):205–214, 2006.

[74] Justin Manweiler, Naveen Santhapuri, Romit Roy Choudhury, and Srihari Nelakuditi. Predicting length of stay at wifi hotspots. In *2013 Proceedings IEEE INFOCOM*, pages 3102–3110. IEEE, 2013.

[75] Ryan Melfi, Ben Rosenblum, Bruce Nordman, and Ken Christensen. Measuring building occupancy using existing network infrastructure. In *2011 International Green Computing Conference and Workshops*, pages 1–8. IEEE, 2011.

[76] Iresha Pasquel Mohottige and Tim Moors. Estimating room occupancy in a smart campus using wifi soft sensors. In *2018 IEEE 43rd Conference on Local Computer Networks (LCN)*, pages 191–199. IEEE, 2018.

[77] Marangaze Munhepe Mulhanga, Solange Rito Lima, and Paulo Carvalho. Characterising university wlans within eduroam context. In *Smart Spaces and Next Generation Wired/Wireless Networking*, pages 382–394. Springer, 2011.

[78] MDC Nokia. MDC 2012 - best challenge entries | nokia research center. http://research.nokia.com/page/12362, June 2012.

[79] Anastasios Noulas, Salvatore Scellato, Neal Lathia, and Cecilia Mascolo. Mining user mobility features for next place prediction in location-based services. In *2012 IEEE 12th International Conference on Data Mining*, pages 1038–1043. IEEE, 2012.

[80] Rajesh Pampapathi, Boris Mirkin, and Mark Levene. A suffix tree approach to anti-spam email filtering. *Machine Learning*, 65(1):309338, 2006.

[81] Bei Pan, Yu Zheng, David Wilkie, and Cyrus Shahabi. Crowd sensing of traffic anomalies based on human mobility and social media. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 344–353. ACM, 2013.

[82] Bartłomiej Płaczek. Selective data collection in vehicular networks for traffic control applications. *Transportation Research Part C: Emerging Technologies*, 23:14–28, 2012.

[83] Sayed W Qaiyumi and Daniel Stamate. Reduction in dimensions and clustering using risk and return model. In *21st International Conference on Advanced Information Networking and Applications Workshops (AINAW'07)*, volume 1, pages 373–378. IEEE, 2007.

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

[85] Alessandro E Redondi, Matteo Cesana, Daniel M Weibel, and Emma Fitzgerald. Understanding the wifi usage of university students. In *2016 International Wireless Communications and Mobile Computing Conference (IWCMC)*, pages 44–49. IEEE, 2016.

[86] Ryan Rifkin and Aldebaro Klautau. In defense of one-vs-all classification. *The Journal of Machine Learning Research*, 5:101–141, 2004.

[87] Salvatore Scellato, Mirco Musolesi, Cecilia Mascolo, Vito Latora, and Andrew T Campbell. Nextplace: A spatio-temporal prediction framework for pervasive systems. In *International Conference on Pervasive Computing*, pages 152–169. Springer, 2011.

[88] Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 253260, 2002.

[89] James Scott, AJ Bernheim Brush, John Krumm, Brian Meyers, Michael Hazas, Stephen Hodges, and Nicolas Villar. PreHeat: controlling home heating using occupancy prediction. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, page 281290, 2011.

[90] Roger W Sinnott. Virtues of the haversine. *Sky and Telescope*, 68:158, 1984.

[91] Peter HA Sneath. The application of computers to taxonomy. *Microbiology*, 17(1):201–226, 1957.

[92] Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818, 2010.

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

[94] Libo Song, Udayan Deshpande, Ulas C Kozat, David Kotz, and Ravi Jain. Predictability of wlan mobility and its effects on bandwidth provisioning. In *Proceedings IEEE INFOCOM 2006. 25TH IEEE International Conference on Computer Communications*, pages 1–13. IEEE, 2006.

[95] H Späth. The minisum location problem for the jaccard metric. *Operations-Research-Spektrum*, 3(2):91–94, 1981.

[96] Helmuth Späth. Cluster analysis algorithms for data reduction and classification of objects. 1980.

[97] Thanchanok Sutjarittham, Hassan Habibi Gharakheili, Salil S Kanhere, and Vijay Sivaraman. Data-driven monitoring and optimization of classroom usage in a smart campus. In *2018 17th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pages 224–229. IEEE, 2018.

[98] Latanya Sweeney. Simple demographics often identify people uniquely. *Health (San Francisco)*, 671:1–34, 2000.

[99] Business Intelligence Analytics Software Tableau. Birkbeck College. `https://cis6.bbk.ac.uk/#/signin`, 2019. [Online; Accessed 18-June-2019].

[100] R Core Team and Others. R: A language and environment for statistical computing. 2013.

[101] Zengshan Tian, Xindi Liu, Mu Zhou, and Kunjie Xu. Mobility tracking by fingerprint-based knn/pf approach in cellular networks. In *2013 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 4570–4575. IEEE, 2013.

[102] Deepak Vasisht, Swarun Kumar, and Dina Katabi. Decimeter-level localization with a single wifi access point. In *13th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 16)*, pages 165–178, 2016.

[103] Long Vu, Quang Do, and Klara Nahrstedt. Jyotish: A novel framework for constructing predictive model of people movement from joint wifi/bluetooth trace. In *2011 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 54–62. IEEE, 2011.

[104] Wen-June Wang. New similarity measures on fuzzy sets and on elements. *Fuzzy Sets and Systems*, 85(3):305–309, 1997.

[105] Klaas Wierenga and Licia Florio. Eduroam: Past, present and future. *Computational Methods in Science and Technology*, 11(2):169–173, 2005.

[106] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.

[107] Jungkeun Yoon, Brian D Noble, Mingyan Liu, and Minkyong Kim. Building realistic mobility models from coarse-grained traces. In *Proceedings of the 4th International Conference on Mobile Systems, Applications and Services*, pages 177–190. ACM, 2006.

[108] Jing Yuan, Yu Zheng, and Xing Xie. Discovering regions of different functions in a city using human mobility and pois. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 186–194. ACM, 2012.

[109] May Yuan. Human dynamics in space and time: A brief history and a view forward. *Transactions in GIS*, 22(4):900–912, 2018.

[110] Mohammed J Zaki, Wagner Meira Jr, and Wagner Meira. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, 2014.

[111] Hui Zang and Jean Bolot. Anonymization of location data does not work: A large-scale measurement study. In *Proceedings of the 17th Annual International Conference on Mobile Computing and Networking*, pages 145–156. ACM, 2011.

[112] Jin Zhang, Bo Wei, Wen Hu, and Salil S Kanhere. Wifi-id: Human identification using wifi signal. In *2016 International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 75–82. IEEE, 2016.

[113] Vincent W Zheng, Yu Zheng, Xing Xie, and Qiang Yang. Collaborative location and activity recommendations with gps history data. In *Proceedings of the 19th International Conference on World Wide Web*, pages 1029–1038. ACM, 2010.

[114] Yu Zheng, Quannan Li, Yukun Chen, Xing Xie, and Wei-Ying Ma. Understanding mobility based on gps data. In *Proceedings of the 10th International Conference on Ubiquitous Computing*, pages 312–321. ACM, 2008.

[115] Yu Zheng, Furui Liu, and Hsun-Ping Hsieh. U-air: When urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1436–1444. ACM, 2013.

[116] Yu Zheng, Yanchi Liu, Jing Yuan, and Xing Xie. Urban computing with taxicabs. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, pages 89–98. ACM, 2011.

[117] Yu Zheng, Lizhu Zhang, Zhengxin Ma, Xing Xie, and Wei-Ying Ma. Recommending friends and locations based on individual location history. *ACM Transactions on the Web (TWEB)*, 5(1):5, 2011.

[118] Changqing Zhou, Dan Frankowski, Pamela Ludford, Shashi Shekhar, and Loren Terveen. Discovering personal gazetteers: An interactive clustering approach. In *Proceedings of the 12th Annual ACM International Workshop on Geographic Information Systems*, pages 266–273. ACM, 2004.

[119] Mengyu Zhou, Minghua Ma, Yangkun Zhang, Kaixin SuiA, Dan Pei, and Thomas Moscibroda. Edum: Classroom education measurements via large-scale wifi networks. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 316–327. ACM, 2016.