Fakultät für Informatik,
Lehrstuhl 11 (Prof. Dr. Schlichter):
Angewandte Informatik / Kooperative Systeme

Context Specific Next Location Prediction

Dipl.-Inform. Halgurt Bapierre

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften (Dr. rer. nat) genehmigten Dissertation.

Vorsitzende:	Univ.-Prof. Bernd Brügge, Ph.D.
Prüfer der Dissertation:	1. Univ.-Prof. Dr. Johann Schlichter
	2. Univ.-Prof. Gudrun Klinker, Ph.D.
	3. Priv.-Doz. Dr. Georg Groh

Die Dissertation wurde am 05.02.2014 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 04.06.2014 angenommen.

Kurz-Zusammenfassung

Die Untersuchung des menschlichen Verhaltens, darunter auch das Mobilitätsverhalten, war lange Zeit auf Umfragen, Experimente mit wenigen Testpersonen oder synthetischen Daten basierend. Die rapiden technologischen Fortschritte der letzten Jahre, insbesondere die Verbreitung von mobilen Geräten wie Smartphones, sowie die Verbreitung des mobilen Zugriffs auf das Internet und die Entstehung von sozialen Netzwerk Platformen erlauben die Erfassung von Unmengen an Informationen über das Verhalten der Nutzer. Neue Dienste sind entstanden, um mobilen Nutzern auf der Grundlage ihrer aktuellen Standorte zu dienen. Die Vorhersage des nächsten Ortes eines mobilen Nutzers erlaubt die Entwicklung von neuartigen, Wertbringenden und intelligenten Dienstleistungen. Dieser Arbeit beschäftigt sich mit der Vorhersage des nächsten Ortes eines mobilen Nutzers. Die Arbeit untersucht Einflüsse von Zeit, Raum, sozialen Beziehungen, anderen Informationsquellen sowie diskreten Wissen (Kalendereinträge) auf das Mobilitätsverhalten des mobilen Nutzers, der Schwerpunkt liegt insbesondere auf soziale Einflüsse.

Short Abstract

The study of human behavior, including the mobility behavior, was based for a long time on surveys, experiments with small numbers, or synthetic data. The rapid technological advances of the last years, especially the pervasiveness of mobile devices such as smartphones, as well as the spread of mobile access to the Internet and the emergence of social networking platforms allow the collection of vast amounts of data containing information about the behavior of users. New services are emerged to serve mobile users based on their current locations. Next location prediction allows the development of more novel, value-making and smarter services. This thesis is concerned with predicting the next location of a mobile user. The work investigates the influences of time, space, social relations, other sources of information and discrete knowledge (calendar entries) on the mobility behavior of a mobile user, where the focus is particularly on social influences.

Contents

Chapter 1

Introduction

Advances in mobile communication, the massive production of powerful yet inexpensive mobile devices (smart phones), the availability of these devices to the general public, the development of mobile communication technologies such as Infrared, Bluetooth, W-LAN, etc. and finally mobile access to the Internet have allowed users to communicate and share information with each other as never before. Furthermore, the development of precise and pervasive indoor (W-LAN, Bluetooth) and outdoor (GSM, GPS) location acquisition technologies and equipping mobile devices with these technologies has fast-tracked the acceptance and pervasiveness of (mobile) social media, facilitating communication between friends and allowing services to be accessed anytime, anywhere.

Location acquisition technologies, pervasive social media, Internet access and sophisticated smart phones help collect a vast amount of data about the current context of the user. Context is everything which surrounds the user and gives meaning to or influence their behavior. The availability of information about the current context of the user has led to the evolution of new kinds of computations such as context-aware and mobile computing. Although the context of a mobile user consists of more than their location, location has nonetheless gained the most attention and in many cases location does in fact stand for the whole context of the user. Context-aware computing has generated new innovative mobile services tailored to the user based on their current location called Location Based Services (LBS) [wik, 2013b]. "Location Based Services (LBSs) are IT services for providing information that has been created, compiled, selected or filtered taking into consideration the current location of the user or those of other persons or mobile objects" [Küpper, 2005]. LBS is the intersection of three technologies, namely Geographical Information Systems (GIS), the Internet and Mobile Services [Vrček et al., 2009]. LBS deals with recreation, entertainment, e-commerce, mobile financial services, location specific mobile advertising, mobile inventory management, wireless business re-engineering, mobile interactive games, traffic information, machine control, etc ([Kang et al., 2009, Varshney, 2001]). The development of new social media has bridged the gap between the physical world and online social networking services and led to the evolution of new services called Location Based Social Networking (LBSN) [mic, 2013]. LBSN allows users to share life experiences, information, activities, interests, etc. anytime, anywhere.

A user's current physical location is central to LBS and LBSN, but knowledge about the next location can be even more important in providing the user with more valuable information and to offer more sophisticated and novel services. Thus predicting the next location of a mobile user given his current context has increasingly gained significance in recent years. Next location prediction can be injected into manifold applications and services. In the following section we present some of the services, that can take great benefits from next location prediction.

1.1 Motivation

Next location prediction can enhance vehicle intelligence (BMW's EfficientDynamics [Eff, 2013]), for example the development of new assistance systems has substantially contributed to safer and more efficient driving [Wevers et al., 2010], optimization of fuel consumption and reduction of co2 emission [Ganti et al., 2010, Ericsson et al., 2006, Lee et al., 2008].

Humans generally drive between only a few significant locations and for such routine journeys do not use their vehicle's navigation system. Next location prediction can be useful in the development of intelligent navigation systems that learn and can then predict the habitual routes of the driver. The navigation system can input valuable information into the onboard systems (BMW's EfficientDynamics [Eff, 2013]) so that braking energy can be recovered not only during the actual downhill [Nav, 2013]. Further, it allows navigation with foresight, informing the driver of upcoming speed limits so that they can reduce their speed gently instead of abruptly [Nav, 2013]. A vehicle can also be prepared for the gear changes required prior to entering a motorway, allowing the momentum to be gained required for the acceleration process [Nav, 2013]. Finally, intelligent navigation systems for suggesting routes aimed at reducing fuel consumption [Ganti et al., 2010, Ericsson et al., 2006, Lee et al., 2008] can benefit from next location prediction so that routes can be recommended prior to the start of a journey, even if the driver does not use their navigation system or enter their destination.

Dynamic Pass Prediction (DPP) is a driver assistance system in BMW's ConnectedDrive service concept for warning the driver of unsafe sections of a route when overtaking [Lee et al., 2008]. The system uses data from several sources including the navigation system to detect unsafe sections. Next location prediction can help to predict the route that the driver is going to take without the navigation system being in use and thus provide the driver with more accurate information about the route sooner (especially when driving in unfamiliar areas).

Intelligent energy management systems such as BMW's ActiveHybrid [Act, 2013] can help reduce energy consumption and emission values by up to 20%. The distance to drive to the next potential destination can be calculated based on predicting the driver's next location. Combining this information with data from the vehicle's onboard systems, topological information about inclines and slopes, etc. allows the system to compile the best energy saving strategy from the moment the drive begins. It determines the best time to switch off the combustion engine and ensure optimum use of the hybrid engine modes. This enables the driver to cover as many kilometers

as possible using the electric motor without producing any emissions.

The performance of a hybrid electric vehicle (HEV) is strongly dependent on the performance of its high-voltage battery pack, which is influenced by the outside temperature. The batteries in HEVs perform badly in cold temperatures. The poor performance of the batteries is caused by high internal resistance which in turn means the vehicle to starts slowly [Pesaran et al., 2003]. Next location prediction can provide the thermal management systems of batteries with travel data before the driver leaves their current location. The thermal management system can then improve the performance of the vehicle by appropriately cooling or preheating the batteries ([Pesaran et al., 2003] as cited by [Krumm and Brush, 2011]).

Computer-aided blogging systems, mobile marketing and intelligent mobile advertising [Barnes and Scornavacca, 2004, Barwise and Strong, 2002] represent other areas where next location prediction can play an important role in providing users with services at a time and location that best suits them. Next location prediction can help both bloggers/advertisers: in order to reduce cost, and subscribers/consumers: in order not to be disturbed at inappropriate times and thereby not losing their interest in the services offered and enabling maximum possible gain from the service. For example, once the time and location of the next activity of the user is determined, the user can be provided with appropriate advertizing related to the activity.

Disaster relief is a another area where next location prediction can play a role ([Gao et al., 2011a, Gao et al., 2011b] as cited by [Gao et al., 2012]). Currently, incidents are reported during a disaster by volunteers and victims who have a communication device (phone, laptop etc). If such people experience limited communication abilities for example in black-out areas, responders, governmental agencies and NGOs face difficulties to get ahead of the demand curve and to be more proactive in deploying aid and rescue capabilities [Gao et al., 2011a]. Next location prediction can combine scientific data about earthquakes, floods, and other phenomena, with crowdsourcing data (user data from volunteers, victims, NGO-employees etc.) in order the location of future requests and needs can be predicted [Gao et al., 2011a]. According to [cro, 2013] crowdsourcing stands for, "the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people and especially from the online community rather than from traditional employees or suppliers" (more detail can be found in [Howe, 2006] as cited by [Gao et al., 2011a]).

Energy saving in private households can also be supported by next location prediction. The USA uses 25% of the total world energy, the residential sector is responsible for 21%, which corresponds to almost 9% of worldwide energy consumption [Gupta et al., 2009]. Experimental results by MIT [Gupta et al., 2009] show that almost 7% of a heating bill can be saved using programmable thermostats that know how far the occupants are away from home ([Gupta et al., 2009] as cited by [Krumm and Brush, 2011]). Next location prediction can help provide a better estimation of the time when an occupant returns home.

In rehabilitation or crime suppression, electronic monitoring and parole is a further use case where location prediction can play an important role. Electronically monitoring people who have committed fairly minor crimes helps reduce costs and ease prison overcrowding, instead of imprisoning them. Further, in the case of hard-

ened criminals, it reduces the risk of corrupting them further and exacerbating the problem by imprisoning them ([Brown et al., 2011] as cited by [Perusco and Michael, 2007]). It can also help protect both victims and offenders in cases involving dangerous criminals, repeated offenders or recidivists on parole. Further, a stalking victim can legally require to be informed of the whereabouts of the offender to avoid unwanted encounters. A next location predictor can help predict the time and location where both victim and offender may potentially meet, either by chance or at the intention of the offender. The victim can avoid the predicted locations and/or the local authorities can increase police presence at them and thus avoid the occurrence of criminal acts.

"A range of applications, from predicting the spread of human and electronic viruses to city planning and resource management in mobile communications, depend on our ability to foresee the whereabouts and mobility of individuals" [Song et al., 2010b]. Traffic management and public transport recommender systems are further examples for the use of next location prediction [Rodriguez-Carrion et al., 2012]. Next location prediction can help provide a user with the traffic conditions on the route to the user's next location.

Combining the location histories of multiple users plus their social network allows the development of friend recommendation systems, the discovery of social trends and generalities, the detection of locations with people sharing the same activities and interests, and last but not least the provision of services like goods sharing (taxi, car, working tools, etc.) [Ye et al., 2009]

Next location prediction can help accurately predict the time when a person is going to be present at certain locations. Presence prediction can be important for example when a person wants to initiate conversations which are sensitive or long, either spoken or typed, only when the other party is at home or in the office, or when the person would rather have a face-to-face meeting [Krumm and Brush, 2011]. Presence prediction could also be useful for an intelligent postal service or when detection of anomalous behavior is important, for example when a disabled person/patient/teenager/child is expected (not) to be at a certain location, but is (not). Further action can be initiated based on this information such as an emergency call [Krumm and Brush, 2011].

Next location prediction can also support assistive technology for disabled or cognitively impaired persons in order to "provide active cognitive aids for people with reduced memory and problem-solving abilities due to Alzheimer's Disease or other disorders" [Patterson et al., 2002]. ACTIVITY COMPASS is an Assisted Cognition system developed by [Patterson et al., 2002] that helps to reduce spatial disorientation inside and outside the home.

The proposed mobility model in this thesis can easily be adapted to a mobility pattern detection system. The focus in these cases is on continuous recording of human movements and pattern detection, rather than on predicting their next location locations. Healthcare monitoring systems can benefit from a system that can detect the mobility patterns of a user. Some monitoring systems are designed to monitor the patient continuously for long-term diagnostic purposes. It is generally inconvenient and inaccurate for a patient to record their activities and behavior. The monitoring system can collect a huge amount of data about the patient using different sensors

including sensors for location acquisition. A mobility pattern detection system can learn the mobility behavior of the user and detect significant behavioral changes, which could in turn could identify causes for changes in the medical condition of the patient [Kim and Kotz, 2011].

1.2 Privacy Issues

Despite the usefulness and convenience of next location prediction specifically and LBS generally, threats to the sphere of personal privacy remain an important and critical issue regarding user acceptance of next location prediction and LBS as well [Shin et al., 2012, Perusco and Michael, 2007, ACLU, 2010, Poolsappasit and Ray, 2009]. According to [Perusco and Michael, 2007] the acceptance of people using LBS is strongly influenced by some of the following factors, "Accountability for the accuracy and availability of location information, prioritization and location frequency reporting, the user's freedom to opt-in and opt-out of services, caregiver and guardian rights and responsibilities, the transparency of transactions, and the duration of location information storage" [Perusco and Michael, 2007]. A service gains a higher acceptance if users have the choice to opt-in or opt-out of a service, if they trust the service provider and their security system, if they know who has access to their location information and with whom they share their information, if they think that the benefits of the service outweigh the risk of invading the individual's privacy, if they have the choice between consent and refusal of monitoring and if the use of the service is trust-building rather than trust-destroying.

The type of service is therefore important in increasing user acceptance of the use of LBS applications involving next location prediction. Generally LBS applications can be categorized according to the type of use into three main categories, namely mandatory, voluntary and non-user applications [Perusco and Michael, 2007]. Examples of mandatory user applications are electronic monitoring prisoners on parole and tracking children and minors by the parents. Examples of voluntary applications are route, navigation, vehicular and domestic applications (intelligent thermostats). Examples of non-user applications are city planning, traffic management and epidemic modeling (spread of human and electronic viruses).

Mandatory user applications are enforced by special laws such as caregiver and guardian rights, etc. providing the police, security authorities, parents the legal rights to keep an eye on criminals or minors, or laws for protection against terror attacks or crime suppression; in line with the belief that the security of the whole of society is more important than the privacy of an individual. The users of voluntary applications have control of their own data and have the choice of using it exclusively for their own use or sharing it with persons they trust. Further, they can opt out or deactivate the service whenever they want. Finally user acceptance of a service can be enhanced if both data and computations are held by the client, for example by the onboard computers of a vehicle or the mobile device of the user.

Sociologically, trust plays an important role in social systems [wik, 2013c]. Trust is attributable to relationships between social actors [wik, 2013c] and since industrialization has also been a feature of the relationship between social actors and

technological achievements. Humans have a natural disposition to trust and to judge trustworthiness that can be traced to the neurobiological structure and activity of a human brain [wik, 2013c]. Humans are therefore more inclined to trust than distrust when they are confronted with people/technologies/objects with which they have not had any (negative) experience. Our natural disposition to trust (also in technology) can be substantiated by the sheer number of users on (location based) social networking platforms which share information on the servers of the service provider. The shared information contains messages, tweets, photos, videos and also information regarding their movements.

In the case of non-user LBS applications, a variety of algorithms have been developed to anonymize the user who requests an LBS service. These algorithms guarantee privacy protection for the users of LBS applications. k-Anonymity [Gruteser and Grunwald, 2003, Kalnis et al., 2007, Gedik and Liu, 2005], k-anonymity without cloaking region [Gong et al., 2010] and SpaceTwist [Yiu et al., 2008] are examples of privacy protection algorithms. The basic idea behind all these algorithms is the anonymization of both the user and their current location. Users send a request to a third party server, which they trust, called the anonymizer. The anonymizer manipulates the user request prior to sending it to the service provider in a manner so that neither the user nor their location can be determined by an attacker. For example, instead of sending the real location of the user a region called K-ASR (k-anonymizing spatial region) is sent to the service provider, where $k-1$ users other than the requester are currently present. The anonymizer filters the best results from the response of the service provider before sending it to the requester.

This thesis was motivated by the usefulness and convenience of next location prediction in providing humans with important benefits. Further, privacy protection is exceptionally important and consequently understanding and studying next location prediction is equally important so that the best protection can be provided. We believe that the work presented in this thesis can provide valuable gain in both cases. Finally, it has been shown by [Song et al., 2010b] that human mobility is predictable to a very high degree, which is a promising result. The points above summarize our motivation to investigate the problem of next location prediction. The focus of this work was not on providing the most accurate mobility model, but rather on showing how the next location of a mobile user can be predicted based on a simple but effective algorithm and how accuracy can be improved by combining spatial with temporal and social features and information from additional sources.

1.3 Research Question

Human behavior has been a subject of scientific interest for a long time. Early research relied on data from surveys (questionnaires) or simulations (synthetic). Technological advances and the development of wearable devices has allowed experiments with several hundred test users to be conducted and made it possible to record data about the context of the test users. However, experimental data is expensive as well as limited because setting up a test is complex and expensive, the test equipment is expensive, finding enough representative test users for the application domain (test persons in most experiments were students and campus employees) is fraught with

difficult and finally, the duration of test is limited (period of data collection). The development of powerful mobile phones such as the smart phone, the integration of a series of sensors in these devices has permitted a huge amount of rich data about the current context of the mobile user data to be collected continuously. Context-awareness [Dey, 2001] is a term from ubiquitous or pervasive computing [Schmidt, 2002] referring to linking contextual data with computer systems [wik, 2013a]. Contextual data allows patterns in the behavior of the user to be learned and once these patterns have been learned, allows future behavior to be inferred based on the current context of the user. Contextual data contains information about current location, time, other users, current device settings, current program usage, etc. In this thesis we refer to the set of spatial features with the spatial context, the set of temporal features with the temporal context, the set of social features with the social context and to other features with additional context.

Analyzing and recognizing pattern in such a vast amount of data is challenging. Machine learning provides a number of methods for pattern recognition. One of the well-studied class of machine learning methods is the Bayesian network. Dynamic Bayesian network (DBN) approaches are based on a method that models sequences of observations of an arbitrary size. Each observation o_t corresponds to a time step t and contains contextual information of arbitrary dimensions, i.e. a set of features such as position, time, social situation, etc. DBN approaches are a general probabilistic approach for estimating the probability of the occurrence of an event/behavior at a future time step given the observations up to the current time step. It is almost impossible to use all past observations to the current time step to predict the future behavior. Further, future behavior/events depend primarily on the most recent observations, therefore at each time step the number of observations is usually limited to the n most recent, n represents the order (also degree of memory) of the model. The limitation of the dependency of the model the n recent observations is known as the Markov assumption.

The DBN approach can capture the conditional dependence between the variables and can solve inference problems under uncertainty. DBN approaches usually consist of two phases, namely the training phase and the testing phase. During the training phase the model learns a set of patterns of length $n+1$ equal to the order of the model plus one. During the test phase the DBN approach takes the last n observations that represent the current context and assigns a probability mass to each of the patterns learned during the training phase. The pattern with the highest probability represent the output of the model, i.e. the predicted event/behavior.

A mobility model of order $n = 2$ can use the sequence of locations visited by the user to learn simple spatial patterns of the form "home - work - home". In order to detect more complex patterns like "home - campus - auditorium - canteen - fitness studio - super market - home" a higher order mobility model is required. The mobility patterns of a user differ in their sizes, mobility models that use a fixed order fail to detect patterns of variable length. The first research question was:

Q1: Can a mobility model be built which predicts the future location of the user based on a context of variable length?

A mobility model of a higher order increases the need for training data. Training DBN approaches is a critical issue and must be well-considered. The performance

of a mobility model based on a DBN approach to a very high degree depends on there being sufficient training data. Sufficient training data means, given the current observation, a probability mass can be assigned to each of the possible patterns (Σ^n, where Σ represents the state space of the model). Unfortunately this condition cannot be sustained, cold-start and zero-frequency problems are known drawbacks of DBN approaches and are the direct consequence of insufficient training data. The findings of this thesis contribute to solving the problem of:

Q2: How can the impacts of insufficient training data be alleviated, i.e. how can the drawback of the cold-start and zero-frequency problems be alleviated?

Context is not limited to location [Schmidt et al., 1999], additional features can lead to improving prediction accuracy. Human mobility exhibits temporal regularities such as Tom goes home every evening (a pure temporal pattern), or whenever Tom visits the campus, he goes to the cafeteria during lunch time (a mixed spatial-temporal pattern), etc. Another major contribution of this thesis is to find answers for the following questions:

Q3: The mobility of humans obeys temporal pattern, how can temporal features be integrated in to a mobility model in order to model the temporal dependency of human movement?

Additional features lead to detecting more sophisticated patterns, but unfortunately adding new features to the model increases the need for more training data. As already mentioned earlier, training data is expensive and sufficient training data is seldom available. A mobility model that is able to detect patterns of variable length can consider more features of the dataset without increasing the need for more training data. In contrast, incorporating more features in a model with variable order helps detect more pattern in the same amount of training data and thus improve prediction accuracy.

Q4: Can a model with variable order improve the prediction accuracy taking temporal features into account without increasing the need for more training data?

Humans are explorative by nature and interested in visiting new locations. Mobility patterns are subject to decay over time. New mobility patterns evolve and older ones decay. The performance of a trained mobility model is sensitive to the changing mobility behavior of the mobile user, especially when change affects very frequently visited locations (for example, a move, new job, marital status or the birth of a child). In many cases the mobility model performs so badly that retraining it is inescapable. A further research question is:

Q5: Mobility patterns are subject to decay, how long can be the life time of a mobility pattern?

Big changes in the life behavior of a user such as a move, a new job, etc. can lead to massive changes in the mobility behavior of the user such as the invalidation of long-term mobility pattern. These changes decrease the performance of the mobility model. A drift function can help to accelerate the decay of invalid patterns. A further research question is the following:

Q6: How can the drift of invalid patterns be accelerated?

Big changes have a further consequence, they lead to the emergence of new mobility patterns. The new mobility patterns require new training data for retraining the mobility model, otherwise the performance of the mobility model decreases massively. A mobility model that is able to simultaneously learn and predict does not necessarily require retraining it. A further question is:

Q7: how can a mobility model be simultaneously trained and tested, i.e. continuously learn new mobility pattern?

The mobility behavior of the user cannot be explained by focusing solely on their own location history. Humans are social animals and subject to social influence. Mobile Homophily is the tendency of two similar individuals to love or to be interested in the same locations. Two questions that arise are:

Q8: How can the similarity between two individuals be quantified? and how can the interest of two individuals in the same locations be quantified?

We refer to similarity between two individuals with social proximity, and their interests in same locations with mobile proximity. We investigate the validity of mobile homophily by answering the following question:

Q9: Does social proximity imply mobile proximity?

Propinquity refers to the tendency of individuals to associate with close things. In accordance with Tobler's first law of geography, "Everything is related to everything else, but near things are more related than distant things" [Tobler, 1970]. Users from the same city influence each other more than users from different cities, users who live in the same building influence each other more than users living in different buildings. Following the idea of propinquity, a further research question is:

Q10: Does physical closure imply higher social proximity?

Human beings have cognitive, emotional, spatial and temporal limits that prevents them from maintaining all their social relationships with the same intensity [Granovetter, 2005]. A user interacts mainly and spends most of their time with a subset of their social relationships, who form together a subgroup with high group cohesion. The emotional needs, beliefs, thoughts, information, locations, times, norms, goals, etc. of users in the same cohesive subgroup overlap to a high extent, which leads to a higher group cohesion. The next research question is:

Q11: How can cohesive subgroups among the social relationships of a user be detected? And how can the cohesion among the members of a subgroup be measured?

Users in the same cohesive subgroup exhibit a high overlap in their information, which may lead to similar (mobility) behavior. The next research question is:

Q12: Do members of the same cohesive subgroup exhibit higher mobility proximity?

Social networks can provide valuable information for detecting the influence of friends on the mobility behavior of a user. Social networks also help detect generalities and social trends that cannot be explained by only taking the individual's data

into account. People are explorative by nature and have a deep-rooted desire for spontaneity. They are interested in visiting and exploring new locations where they have never been before. In many cases new locations are recommended by friends who have visited these locations recently. Thus social networks can contribute to predicting locations where the user has never been before, especially at weekends, evenings and during free time when people are most explorative. We address in this thesis the problem of transmitting social influence into a predictive model to find an answer to the question:

Q13: How can social influences be modeled to improve location prediction based on the impacts of social networks?

The social relationships among members of same cohesive subgroups represent strong ties of the users. The overlap between the emotional needs, information, behavior, etc. of two users connected via a strong tie is considerably higher compared to two arbitrary friends, therefore they exchange only in seldom cases new information. The overlap between the information of members of different cohesive subgroups is considerably lower, therefore, therefore they exchange increasingly more novel information. The next research question is:

Q14: How information flows between different social groups?

Social influence is subject to decay. The amount and duration of information sharing is a critical issue regarding the acceptance of a service. A user may accept sharing their location history with their friends for a limited period of time, but if this time exceeds a certain value the user may feel unhappy and reject sharing. A research question regarding the drift of social influence is:

Q15: Does social influence decay?

and

Q16: How long should location histories of friends be stored and used to improve prediction accuracy?

Today's mobile devices, specially smart phones, are equipped with additional sources of information such as different sensory like WLAN, Bluetooth, accelerometers, etc. various applications like clock alarm, system programs, etc. and mobile access to other information-sources/applocations via the internet. These additional sources of information allow the collection of huge amounts of data about the users carrying these devices. A further research question is:

Q17: Can the inclusion of additional sources of information enhance prediction accuracy?

If yes, then

Q18: How can the model be extended, in order to take new information sources into consideration, which could help detect more sophisticated patterns?

The mobility behavior of human beings consists mainly of regular patterns and a small part of irregular movements. The irregular movements of a user does not follow any pattern, but nevertheless, users leave evidences that may indicate to their irregular movement behavior. An example of such an evidence is a calendar

entry containing the next dentist appointment or the date of the next concert of the favorite band. We refer to this kind of evidences as discrete knowledge. Discrete knowledge differs from contextual knowledge in its validity. Discrete knowledge expires after the point of time of its expected occurrence elapses, whereas contextual knowledge represents the habits of the user that are valid over a longer period of time. A further research question addressed by this thesis was:

Q19: How can the probability of visiting a location based on discrete knowledge be boosted/dampened?

Finally, modeling condition interdependence is another challenge if the DBN approach is used for prediction tasks. In this thesis we also look at feature selection and how to model the conditional interdependence between the selected features.

The research questions mentioned above are investigated using two scientific method approaches, namely design science and empirical evidences. From the point of view of design science we investigate next location prediction by designing a solution, defining the features/variables of the designed solution, implementing the solution and finally evaluating both solution and implementation using empirical datasets, in order to answer some of the research question. From the point of view of empirical evidences we extract models from the empirical datasets using induction in order to answer the remaining research question. We use design science to find answers for the questions (Q1, Q2, Q3, Q6, Q8, Q11, Q13, Q18, Q19), and empirical evidences for answering the questions (Q4, Q5, Q7, Q8, Q9, Q10, Q12, Q14, Q15, Q16, Q17).

We use three experimental datasets collected by test subjects over a limited period of time, and one real life dataset for evaluating the designed solutions and for empirically answering the research questions, namely **GeoLife** [Geo, 2013a], **Reality Mining** [Eagle and Pentland, 2006], **Foursquare** [fou, 2013] and **Mobile Data Challenge (MDC)** [Laurila et al., 2012].

The following mathematical-statistical correlations are used throughout this thesis for emphasizing the answers to the research questions:

- **Pearson's Correlation Coefficient** is a measure for showing any linear statistical dependence between two random variables, i.e. to assess how well the relationship between the two variables can be described using a linear function (for example, humans tend to move more slowly with increasing age). The correlation coefficient has a value in the interval $[-1, 1]$, a value near zero indicates no correlation, a value near -1 or $+1$ indicates a strong negative or positive correlation between the variables.

- **Spearmann's Rank Correlation Coefficient** is a non-parametric measure of the statistical dependence between two random variables which assesses how well the relationship between two variables can be described using a monotonic function. It calculates the correlation coefficient based on the rank of the values of both variables, in other words it is Pearson's correlation coefficient for the ranks of the values. Spearman's correlation coefficient is thus less sensitive to strong outliers and achieves good results for all monotonic functions rather than for linear functions only. Additionally to Spearman's correlation coefficient we calculate the corresponding p-value $P(\epsilon)$. The p-value $P(\epsilon)$ in

statistical significance testing is the probability of the "null-hypothesis", i.e. that two phenomena have no relationship [Goodman, 1999]. Researchers reject the null hypothesis if the p-value is less than a threshold, often 0.05.

- **The Shannon Entropy** quantifies the measure of uncertainty in the probability distribution of a random variable [Russell and Norvig, 2010, Page 703]. A random variable that takes many different values with similar probabilities means that: predicting which value the random variable takes is subject to high uncertainty.
- **Student's T-Test** is a statistical hypothesis test that can determine whether two sets of data are significantly different from each other. We use this method to determine whether the mean of the improvements in accuracy is different from zero using a basic model and an extension (by adding new features to the basic model) for predicting the mobility of the same set of users

1.4 Thesis-Structure & Chapter Summaries

Chapter 2

Human beings are explorative by nature and have a deep-rooted desire for spontaneity as stated earlier. nevertheless, human mobility exhibits high regularity following a few movement pattern which leads to high mobility predictability [Song et al., 2010b]. People usually commute between few significant locations, which they visit in certain orders, for example "home - work - home". We refer to the order of visiting significant locations with spatial pattern and the dependency of human mobility on these spatial patterns with spatial dependency. A mobility model can predict the future location of a mobile user based on their spatial patterns. The location history of a user is needed in order to learn their spatial patterns. The rapid technological developments over the last two decades have allowed access to various sources of information anytime anywhere and allow a huge amount of data to be harvested from these sources of information. Sources of information are, for example, integrated GPS, W-LAN, Bluetooth devices, social networks and social networking platforms such as Twitter [twi, 2013], Facebook [fac, 2013], Foursquare [fou, 2013], etc. These developments allow the collection of huge amount of data about the movements of a mobile user, to which we refer with location history for simplicity reasons. The availability of location history has prompted the evolution of a new class of computing, namely location-aware computing. Location-aware computing builds the foundation for Location Based Services (LBS). This chapter provides an introduction to topics important for LBS like location acquisition, mining location history, mobility models, location prediction etc.

Chapter summary: Section I of chapter 2 focuses on mining location history, pattern recognition, feature selection and related works based on mining location histories of mobile users. Section II provided an overview over a set of different mobility model approaches. Section III provides an overview of location acquisition technologies. The focus of section IV is on detecting significant locations of users based on raw GPS-data. A framework for choosing the correct mobility model, in terms of

design and evaluation criteria and which highlights both the strengths and weaknesses of the model is presented in section V. Section VI provides an introduction to the mobility model approach used in this theses. The last section VII of the chapter is reserved for the empirical results based on several datasets.

Chapter 3

Context is not limited to spatial information. The spatial context helps detect frequent spatial pattern in the location history of the user, the inclusion of temporal context helps detect the returning intervals of a frequent spatial pattern. For example, a user might visit every Tuesday and Thursday evening after leaving their working location a fitness studio, and in the remaining three working days they go home. Thus the spatial pattern home-work-fitness studio has two temporal patterns (returning intervals), namely Tuesday and Thursday evening. Temporal patterns indicate the periodicities within which a user exhibits the same spatial (sub)pattern.

The mobility of humans obeys both temporal and spatial-temporal patterns. A mixed spatial-temporal pattern might be, "Whenever Bob visits the campus, he goes to the neighboring Mediterranean restaurant for lunch at around 12 o'clock." Some patterns are dependent on the temporal context of the user only, such as, "Bob goes home every evening, despite his current location.". Periodicities can be described using temporal features such as the day of the week, hour of day and off-days. Mobility patterns are subject to decay over time. Incorporating the decay of mobility patterns at the design stage reduces the need for retraining the mobility model.

Chapter summary: We present spatial-temporal and pure temporal pattern in section I. Section II introduces research into topics related to periodicities and temporal pattern. Mining periodic pattern and methods for detecting periodicities are the subject of section III. In section IV we introduce our mobility model, which is capable of detecting spatial, temporal or spatial-temporal pattern. The methods addressed in this chapter are evaluated in the last section using three different datasets.

Chapter 4

Homophily refers to the tendency of individuals to associate with similar individuals. Similar individuals have in common features such as interests, beliefs, thoughts, locations, emotional needs, goals, norms, etc. Mobile homophily refers to the tendency of similar individuals to be interested in the same locations, thus it involves two measurements, namely social proximity and mobile proximity. Mobile proximity quantifies the similarity between the mobility of two individuals based on mobility measurements such as common spatial and spatial-temporal locations. Network proximity quantifies the similarity between two individuals of the social network side. Network similarity can be determined using measurements from social network analysis (SNA). The focus of this chapter is on showing statistical dependence (if any) between social and mobile proximity. Due to cognitive, emotional, spatial and temporal limits, people cannot maintain of all their social relationships with

the same intensity [Granovetter, 2005]. Each user can have groups of strong social relationships, in which almost everyone knows everyone else, thus the cohesion among the members of the same group is considerably higher compared to a group of arbitrary selected friends. The correlation between group cohesion and mobile homophily is a further subject of this chapter. The chapter contains methods for analysing both social networks and location histories of different users.

Chapter summary: Section I provides a brief introduction in social networks, possibilities arising through technological achievements and the growing interest in investigating the influence of social networks in real life phenomena in the last two decades. Section II contains a formal description of social networks, their properties and interior structure. Further, the section provides an introduction of various types of locally dense regions representing cohesive subgroups, their properties, methods for their detection and a measurement for calculating group cohesion. Section III provides concepts for calculating social proximity among users. Section IV contains an introduction to the propinquity effect and methods for calculating mobile proximity. Section V presents some relevant works about the interplay between spatial properties such as locations and distance and social networks. Section VI provides an in-depth correlation analysis between social and mobile proximity based on data collected from March to July 2012 from an LBSN platform called Foursquare. The section sets particular focus on the effects of propinquity and cohesive subgroups on the mobility behavior of mobile users.

Chapter 5

The influence of social networks on an individual's behavior has been the subject of much research. The behavior of an individual cannot be explained by focusing solely on the behavior of that individual user. Mobility is, as is any other human behavior, subject to social influences. Human beings are both explorative and social in nature. They share their experiences of exploring the environment with their friends, family members, etc. Sharing experiences may be synchronous, for example when two friends visit the same restaurant, or general social trends in the community of the user like a trendy shop, a hip night club or other locations implicitly recommended by friends. The focus of this chapter is on the causation effect between social proximity and the mobility behavior of socially connected individuals. Causation effect means whether network similarity causes similar mobility behavior among different individuals. We show the causation effect by using the mobility model of an individual with social network influences to improve the accuracy of predicting the next location of the user.

Chapter summary: Section I is a general introduction to the causation effects of social networks on human mobility. Section II provides an overview of the existing social influence models. Section III focuses on extending the general spatial-temporal model of chapter 3 for integrating influences from the social network into the individual mobility model. The last section IV contains empirical conclusions about the causation effect between network similarity and next location prediction based on a dataset from the location based social network platform Foursquare.

Chapter 6

Context is not limited to spatial, temporal and social features. Mobile devices such as smart phones contain a set of sensors that can log a huge volume of data about many aspects of an individual's behavior. Logs of program usage, Bluetooth, W-LAN and system settings are examples of additional sources of information that could help detect more patterns in the individual mobility of a mobile user. A further type of information is knowledge that is valid during a certain period of time and then expires later, such as calendar entries about appointments, cinema or restaurant reservations, etc. We refer to this information as discrete knowledge. Discrete knowledge helps boost and dampen the probability of visiting certain locations and accordingly improves the accuracy of the mobility model. This chapter focuses on integrating additional information sources and discrete knowledge into the individual mobility model with the aim of increasing the accuracy of next location prediction.

Chapter summary: Section I provides an introduction to additional information sources, the integration of additional information sources into the mobility model, related work, the MDC dataset and empirical results. Discrete knowledge, people's adherence to schedule, the extension of the mobility model in order to incorporate discrete knowledge and finally empirical results are presented in section II.

The final chapter of the thesis contains a summary of the conclusions followed by critical discussion and future prospects.

Chapter 2

Location Prediction Bases for Spatial Dependency

The development of inexpensive mobile devices, wireless networking and the availability of precise location acquisition technology has facilitated the collection of huge sequences of precise data about the users carrying these devices. We refer to sequences of such data with movement or location history. Research into human behavior has been hampered for a long time by the lack of (sufficient) real life behavioral data. Consequently, the investigation of human behavior was based on surveys and synthetic data, which was far removed from realistic human behavior. The availability of the location histories of mobile users has opened up new opportunities for pervasive investigation and the understanding of human behavior at anytime and anywhere. The availability of location history has prompted the evolution of a new class of computing, namely location-aware computing [Lu and Liu, 2012, Patterson et al., 2003a]. Location-aware computing is the foundation for location based services, where the geographical location of a mobile user is used to tailor to the user a wide spectrum of new (personalized) services.

Mobile users exhibit regularity in their movements that follows certain patterns. These patterns make the movement of mobile users predictable to a high degree [Gonzalez et al., 2008]. The prediction of the future location of a mobile user based on their current location and/or n previous locations forms the framework for the spatial dependency of human movement and is the subject of this chapter.

Chapter summary: Section I provides an introduction to location history mining, pattern recognition, feature selection and related works based on mining location histories of mobile users. An overview of existing mobility models and their functionalities is presented in section II. An overview of location acquisition technologies is presented in section III. Location representations and methods of extracting the user's significant locations based on their raw movement traces are presented in section IV. Section V provides a set of general requirements for the choice of a proper mobility model, as well as a set of design and evaluation criteria that are important for evaluating it. The mobility model used in this thesis is presented in section VI. The last section VII of the chapter is reserved for the empirical results based on several datasets.

2.1 Location History Mining

Location history is a sequence of time stamped location based observations collected by different sources of information. Sources of information are, for example, integrated GPS, W-LAN, Bluetooth devices, social networks and social networking platforms such as Twitter [twi, 2013], Facebook [fac, 2013], Foursquare [fou, 2013], etc. We refer to the collected values at each time step as the observation e_t at time step t. Each observation contains a set of features such as location data collected by a GPS-Sensor, other mobile devices seen in the neighborhood by the Bluetooth device, W-LAN access points sensed by the W-LAN device, different application entries/logs such as calendar, alarm clock, system profile and the call-log of the mobile phone, etc. thus an observation can be of arbitrary dimensions, containing evidence about the spatial, temporal, social and other contexts of the user.

The daily life of a user has a tight relationship to the geographical locations they visit, thus the location history of a mobile user contains lots of valuable information that helps in understanding their life style [Ye et al., 2009]. Location history mining aids the extraction of higher level knowledge such as regularities, anomalies or phenomena from sequences of low level sensor data, for example recognizing regular movement patterns or significant locations from the location history of a user. Pattern recognition is thus an important task for location history mining.

2.1.1 Pattern Recognition

The power of our intelligence lies in our ability to observe/experience our surroundings through our senses and to comprehend our environment based on those observations/experiences, combined with our cognitive capabilities to make/choose correct decisions/actions. Pattern recognition deals with the identification of valid patterns of an arbitrary size, shape and complexity using sequential data as input and performing an action according to the category of the pattern [Duda et al., 2001].

"The ease with which we recognize a face, understand spoken words, read handwritten characters, identify our car keys in our pocket by feel, and decide whether an apple is ripe by its smell belies the astoundingly complex process that underlies these acts of pattern recognition" [Duda et al., 2001].

The detection of pattern in the raw location history is highly dependent on the contained features. "A feature is an individual measurable heuristic property of a phenomenon being observed" [fea, 2013]. A simple pattern like "home - work - home" contains one spatial feature, namely locations, whereas a more complex pattern like "whenever Bob meets Tom at the campus, they go for lunch to the neighboring pizzeria" contains spatial, temporal and social features. The selection of proper features is key to any successful pattern recognition algorithm.

2.1.2 Feature Selection

The observations may contain redundant or irrelevant features for pattern recognition, redundant features occur for example when two sources of information such as

GPS sensor and WLAN device identify locations. An irrelevant feature for mobility might be the settings of the browser for example. Further, some features may be dependent on other features, for example bluetooth observations are relevant only when the carriers of both devices have a social relationship. Thus, feature selection (also known as variable or attribute selection) is an important task of pattern recognition. Feature selection deals with identifying the subset of features that influence the construction of the (predictive) model, eliminating redundant or irrelevant, and identifying interdependence among the features. Feature selection affects both model performance and training time, thus it is an important task in model construction. The following list provides four different techniques for feature selection [Dietterich, 2002]:

- **The wrapper approach** has two variants, namely forward selection: constructing the model with one feature and successively adding more features, and backward selection: adding all features to the model and successively eliminating features. Each time a feature is added (or removed) the performance of the model must be evaluated in order to select the subset of features that maximizes the performance of the model.
- **Penalty placing techniques** include all the features in the model and place penalties on the parameters associated with the features. The idea is that the impact of parameters associated with useless features becomes very small or even perhaps reduces to zero.
- **Measure of relevance** computes some measure of feature relevance and removes the features with a score under a certain threshold.
- **Simple model fitting** first fits a simple model and then analyzes the fitted model to identify the relevant features and removes the features with a low influence on the model.

In this thesis we follow the wrapper approach starting with one feature and successively extending it by adding more features under consideration of the interdependency between the selected features and measuring the improvement in accuracy. We refer to features relating to location measurements with spatial context. Temporal context is used for features relating to time. Features from social networks build the social context. We refer to additional features with additional context.

2.1.3 Related Work

Mining sequential data generated by a mobile user has been the subject of much research. In the next few subsections we present a set of related research on topics that are closely associated with mining sequential data.

2.1.3.1 Significant Location Detection

The goal of significant location detection is the discovery of a user's specific important locations [Ashbrook and Starner, 2003, Kang et al., 2004, Liao et al., 2007a].

[Ashbrook and Starner, 2003] makes use of both raw GPS measurements and the time a user spends at a location to find the significant locations of that user. All GPS observations within a radius r and with a stay time greater than a certain threshold belong to the same significant location.

[Kang et al., 2004] has proposed an algorithm that allows online detection of significant locations. The clustering algorithm is similar to the k-mean variant used by [Ashbrook and Starner, 2003] with an additional stay time threshold for controlling the amount of time a user spends in a cluster. Consecutive data points form a cluster if they all lie within a radius r and the time elapsed between the first and the last data point exceeds a threshold. Consecutive data points during movement do not form a cluster, either because they lie within a distance greater than r or the time elapsed between the first and the last data points is less than the time threshold.

[Liao et al., 2007a] has proposed a hierarchical model for detecting significant locations as follows. The first level of the hierarchy is formed by the raw GPS traces. Consecutive GPS data points are spatially segmented to discrete spatial regions of radius r. The activity of the user is determined by making use of the duration between the first and the last data point within a segment, the time of the day and the availability of additional geographical information about restaurants, stores, bus stops, etc. in that region. Once the activity of the user is determined, a significant location where the user can perform that activity can be estimated in the next level of the hierarchy.

2.1.3.2 Activity Recognition

[Eagle and Pentland, 2009] infers the daily routines of a user via principal component analysis (PCA) using mobile phone sensor data and GSM traces collected by a mobile phone provider. An individual's daily behavior is modeled as a vector $n \times 24$ with n labels corresponding to behavior and the 24 hours of a day. Given the daily behavior of a user over a period of D days, the model can extract the most prevalent behavior of that user over the hours of the day. Given the prevalent behavior of the user and the behavior of the user over the first half of the day, the model can predict the user's behavior over the remaining 12 hours of the day with an accuracy of 79%.

[Liao et al., 2007a] estimates the activity of a mobile user given his GPS traces. The model depends heavily on temporal features and the existence of a geographical dataset containing locations (such as restaurants, shopping malls, etc.) in the vicinity of the user's current location.

2.1.3.3 Mobility & Motion Detection

[Krumm and Horvitz, 2004] has proposed a model for learning and inferring the motion mode (whether the user is moving or staying) and the location of a user in an indoor environment. The proposed model is called LOCADIO [Krumm and Horvitz, 2004]. It uses Wi-Fi signal strengths from existing Wi-Fi access points in a building complex and mobile Wi-Fi receivers carried by the users. The model is based on the assumption that the signals received by a device are nosier when the user moves. Once the motion mode is predicted, the next location of the user can

be predicted taking into account path constraints (such as walls and doors) as well as human pedestrian speed. Both motion mode and location predictors are based on Hidden Markov Models (HMM) [Krumm and Horvitz, 2004].

2.1.3.4 Transportation Mean

Given the location and velocity of a user, the location of their car and the current observations from a wearable GPS device, [Liao et al., 2007b] has proposed a model to simultaneously learn and infer both the user's transportation mode such as car, foot or bus and their destination location [Liao et al., 2007b, Patterson et al., 2003b, LIAO et al., 2006]. The model can learn the locations where the user changes their transportation mode as well as their destination from the raw GPS data in an unsupervised manner. The proposed model is based on an abstract hierarchical Markov model, more specifically a Rao-Blackwellized particle filter is applied by the authors to make more efficient inference from both raw GPS data and higher level knowledge extracted from the raw data. The inference of both transportation mode and location can be used to provide the user with valuable information such as information about the traffic conditions in a certain area or the current timetable at a bus stop [LIAO et al., 2006]. The model can also be used to detect anomalous behavior such as getting on the wrong bus [Patterson et al., 2003b], which can be used in turn to initiate proactive alerts or calls for assistance in order to help cognitively impaired users through their daily life [Patterson et al., 2004].

2.1.3.5 Location Based Social Networking

Social networks and human mobility influence each other mutually. Users who share information, thoughts, any kind of objects, geographical areas, etc. exhibit similarities in their behavior. The calculation of similarity between two users helps to infer their behavior and make recommendations to a user based on data of another user. In [Han et al., 2007, Zheng et al., 2009a] and [Li et al., 2008] the location histories of users are used to calculated a location based on the similarity of two users. [Li et al., 2008] has proposed a framework to infer the similarity of two users based on the geographical regions shared between them. The more locations two users share, the more similar they are. A similar work in [Zheng et al., 2009a] uses the GeoLife dataset 2.7.1 for calculating similarities between users. The entire social network could be reconstructed based on location histories of users in [Geo, 2013a].

An in-depth correlation analysis between social proximity and mobile homophily has been made by [Wang et al., 2011]. The experimental results do indeed show the statistical dependence between measurements relating to social proximity and mobile homophily. [Scellato et al., 2011c] has investigated the link prediction problem based on location histories of users. The experimental results showed that future friendships between users who share the same locations, but who are not yet friends can be predicted. The calculation of tie strength between users who are socially connected based on their location histories has been investigated in [Wang et al., 2011]. [Eagle et al., 2007] has investigated inference of the structure of social networks using a dataset collected by a mobile phone provider.

2.1.3.6 Location Prediction

Location prediction is one of the main research areas based on the location histories of users. A large number of mobility models have been developed within the last decades. The mobility models differ in their applicability to the spatial environment. Generally, two spatial environments can be distinguished, namely local/indoor [Tran et al., 2008, Kaemarungsi, 2005, Kaemarungsi and Krishnamurthy, 2004, Petzold et al., 2005] and global/outdoor [Mathew et al., 2012, Ashbrook and Starner, 2003, Eagle and Pentland, 2009]. Different types of location acquisition technology can be applied depending on the spatial environment, i.e. area maps, GPS, GSM or geographical databases. Different location acquisition technologies vary in their representation of location data such as continuous (GPS), cellular (GSM) and discrete (Points of Interests (POI), building/areal maps, etc.)

Over the course of the last few decades, different mobility model approaches have been developed such as the random mobility model [Bettstetter, 2001b, Bettstetter, 2001a], the linear dynamical system (LDS) [Liang and Haas, 2003], fixed order Markov model (FOMM) [Ashbrook and Starner, 2003], Neural Networks (NN) [Vintan et al., 2004], Dynamic Bayesian Network (DBN) [Petzold et al., 2005, Krumm and Horvitz, 2006, LIAO et al., 2006], Hybrid Markov Model [Yu et al., 2006], Hidden Markov Model (HMM) [Mathew et al., 2012, Krumm and Horvitz, 2004], Principal Component Analysis (PCA) [Eagle and Pentland, 2009] and the Variable Order Markov Model (VOMM) [Begleiter et al., 2004, Bapierre et al., 2011].

2.2 Mobility Models

The prediction of the next location of a mobile entity generally, and a mobile user specifically, has been the subject of many works. Mobility models represent a set of well-studied algorithms for location prediction tasks. The mobility models vary according to the model assumptions and the input/output parameters they use. The models can be grouped into different categories. There are random mobility models where the model parameters are chosen randomly, models of exact inference, i.e. analytical models where the exact location of a mobile entity is calculated based on the laws of physics, and models of approximate inference based on probabilistic reasoning, i.e. assigning a probability mass to each location and predicting the location with the highest probability.

2.2.1 Random Mobility Models (Traceless)

Random mobility models do not depend on the location history of mobile entities. The parameters for the model such as speed, direction and destination are chosen randomly without taking into consideration restrictions like obstacles, geographical conditions and other entities. Random mobility models are usually used for simulation purposes in mobile ad hoc networks (MANET). Random mobility models try to calculate the position of a mobile entity analytically, based on few simple assumptions such as which of the model's parameters (speed, direction and destination) should be chosen randomly, when to change/replace the chosen parameters

and what should happen when the mobile entity reaches the border of the simulation area [Bettstetter, 2001b].

- **Random walk -** in this model at each time step a mobile entity is assumed to move at a random speed chosen from the range $[0, V_{max}]$ and in a random direction chosen from the range $[0, 2\pi]$ ([Zonoozi and Dassanayake, 2006] restricts direction changes to the upper bound $\Delta\theta_{max}$). The mobile entity continues moving with the chosen speed and in the chosen direction until it reaches the boundary of the simulation area. Upon reaching the boundary, the mobile entity is bounced back into the simulation area at an angle $\theta(t)$ or $\pi - \theta(t)$ [Hong and Rappaport Stephen, 1986]. The random walk model does not allow the entities to spend any time upon arrival at their destinations.

- **Random way point -** the random way point mobility model takes into consideration pause times. The mobile entity moves to a pre-selected destination (instead of moving to the boundary of the simulation area). Upon reaching the destination, the mobile entity pauses for a certain pause time τ_{pause} [Johnson and Maltz, 1996]. Setting τ_{pause} to zero equates to continuous mobility. The density of mobile entities near the center of the simulation area becomes very high a while after the simulation has started and almost zero near the boundaries.

- **Boundless simulation area -** at each time step t, both speed v_t and direction θ_t are calculated based on the speed and direction at the previous time step $t-1$. The speed is calculated considering the acceleration a_t which is chosen from the range $[-a_{max}, a_{max}]$. The direction is calculated by adding a randomly chosen direction change $\Delta\theta_t$ to the previous direction θ_{t-1} from a range $[-\Delta\theta, \Delta\theta]$. Instead of bouncing the entities when the simulation area is reached, the boundless simulation area allows the mobile entities to cross the boundary, reappear at the opposite side of the simulation area and continue travelling with the same speed and in the same direction [Haas, 1997].

- **Smooth random mobility**- in the previous models, both speed and direction are chosen independently from their previous values (except a rough consideration in Boundless simulation area), which causes unrealistic movements such as sharp turns and sudden acceleration/deceleration. The smooth random mobility model considers both speed and direction changes to be smooth. The speed changes incrementally, taking current acceleration into account. Change in direction occurs at several time steps [Bettstetter, 2001b].

- **City section -** the previous models do not consider geographical restrictions such as the topology of the area. [Bai et al., 2003] proposed a work, in which the movements of mobile entities are restricted by the street map of a city section and its speed limits.

- **Group mobility models -** the previous models assume an entity moves freely without being restricted by other entities. [Hong et al., 1999, Sánchez and Manzoni, 2001] introduced many group mobility models where the movement of an entity is additionally restricted by the movement of other entities. The Pursue Mobility Model, Column Mobility Model, Reference Point

Group Mobility Model, Exponential Correlated Random Mobility Model and the Nomadic Community Mobility Model are a few examples of group mobility models.

- **Obstacle Mobility Model** - this mobility model has been proposed by [Jardosh et al., 2003] and takes geographic restrictions such as buildings into account while modeling the movement of mobile entities. The authors use Voronoi diagrams to model buildings and paths between them.

2.2.2 Linear Dynamical Systems (LDS)

Linear Dynamical Systems have a set of mathematical properties and can be solved exactly. LDS assume a linear relationship between both observations and states vectors with zero mean and additive Gaussian noise [Krumm et al., 2003, Bishop, 2007, P 636-637]. The Kalman filter is an example of an algorithm for handling linear, continuous variables. LDS consists of two steps, namely the prediction and update steps.

$$X_t = A_{t-1}X_{t-1} + \underbrace{B_{t-1}u_{t-1}}_{\text{controlled input}} + \underbrace{a_{t-1}}_{\text{random noise}} \tag{2.1}$$

Equation 2.1 represents the prediction step, where A_{t-1} is the transition matrix, x_t is the state at time step t, u_t is a controlled input representing the deterministic part of the noise with its dynamics captured in the matrix B_t and finally a_t represents the uncontrolled random noise from the distributions $a \sim \mathcal{N}(a|0, \alpha)$. Both u_t and a_t have Gaussian distributions with a zero mean [Krumm et al., 2003].

$$E_t = HX_t + w_t \tag{2.2}$$

Equation 2.2 represents the correction steps, where the relationship between both observation E_t and state X_t is linear Gaussian with additive Gaussian noise w_t with a zero mean. H is the measurement matrix and w is chosen from the distribution $w \sim \mathcal{N}(w|0, \omega)$ [Bishop, 2007, P 636-637]. For example, we may set $X_t = V_t, E_t = S_t$, interpreting S_t as actual locations and V_t as speed and choose the model parameters B, H in a manner such that S_t and V_t are linearly dependent plus added Gaussian noise $P(X_t|X_{t-1}) \sim \mathcal{N}(S_{t-1} + \Delta V_{t-1}, B)$ where Δ is the time difference between the discrete time steps t and $t-1$.

A mobility model based on LDS has been proposed by [Liang and Haas, 2003]. The linear relationship between the state and measurement variable is adjusted using a parameter to control the degree of memory of the model α. The state of the model X_t consists of two variables $X_t = (V_t, D_t)$ for estimating both velocity V_t and direction of movement D_t at time step t. μ_V and μ_D are both asymptotic, which means t approaches infinity for both velocity and direction respectively. a_t and θ_t represent the acceleration and direction change respectively and are independent, uncorrelated and Gaussian processes with zero means and standard deviations σ_a and σ_θ [Liang and Haas, 2003]. Both equations 2.3 and 2.4 represent the prediction steps for both

velocity and direction of movement (for simplicity, a first order Markov and a time slice equal to one is assumed).

$$V_t = \alpha V_{t-1} + \underbrace{(1-\alpha)\mu_V}_{\text{controlled input}} + \underbrace{\sigma_a\sqrt{(1-\alpha^2}a_t}_{\text{random noise}} \tag{2.3}$$

$$D_t = \alpha D_{t-1} + \underbrace{(1-\alpha)\mu_D}_{\text{controlled input}} + \underbrace{\sigma_\theta\sqrt{(1-\alpha^2}\theta_t}_{\text{random noise}} \tag{2.4}$$

Once the current states variables V_t and D_t of the model are predicted then the exact location l_{t+1} of the mobile user can be predicted according to equation 2.5

$$l_{t+1} = l_t + V_t \cos D_t \tag{2.5}$$

Setting α to zero means a totally random process whereas a linear motion is obtained by setting α to one [Liang and Haas, 2003].

Although LDS models can be used for the short term (a few seconds to a few minutes) mobility predictions based on the reasoning of [Liao et al., 2007a, Liang and Haas, 2003] (by discretizing the street map to 10 meter cells), the assumptions concerning Gaussian distributions and fixed Markov order are usually not applicable to human motion on an intermediate timescale of several hours or even days since human trajectories show a high degree of spatial and temporal regularity [Gonzalez et al., 2008] which is not well-captured by the aforementioned random models. Other studies have also confirmed the highly non-random nature of human mobility [Song et al., 2010a]. Furthermore, the linear system dynamics make the detection of sharp turns and the avoidance of motion through walls difficult [Krumm et al., 2003].

2.2.3 Mobility Models Based on Probabilistic Reasoning

Mobility models based on probabilistic reasoning predict the future location of a mobile user using sequences of past observations (location history) [Bishop, 2007, P 606]. A mobility model, which can depict realistic human movements, must be able to extract as many patterns as possible from the observations. The sequences of observations are usually noisy and contain uncertainty. Examples of uncertainty are errors in location measurements such as acceleration/deceleration while measuring the velocity of a mobile user or sensor errors like measurement errors in GPS coordinates. The uncertainty increases rapidly when predicting the location of the user far into the future (hours and days). Mobility models based on probabilistic reasoning are more attractive because of their ability to handle uncertainty. They use a belief network to calculate the probability distribution of the observations instead of calculating the exact location of a user [Krumm et al., 2003].

Equation 2.6 demonstrates the (exact) inference problem using the full joint distribution over the complete observation history up to time t:

$$p(e_1, e_2, ..., e_{t-1}, e_t) = \prod_{n=1}^{N} p(e_t|e_1, ..., e_{t-1}). \tag{2.6}$$

2.2.3.1 Fixed Order Markov Model (FOMM)

The size of the observation history increases as t increases. The exact inference become very complex because t is unbounded, thus a mobility model based on the full joint distribution is impractical. Fixed order Markov models estimate the joint distribution approximately based on the Markov assumption. The Markov assumption implies that each observation is independent from previous observations except from a finite set containing the n most recent observations. This means that consecutive observations are highly correlated and future observations older than n time steps are decorrelated, i.e. the effect of the whole of the past observations on e_t is included in the n preceding observations [Bishop, 2007, P 606]. Equation 2.7 indicates the joint distribution of an order $n = 2$ Markov model

$$p(e_1, ..., e_N) = p(e_1)p(e_2|e_1) \prod_{n=3}^{N} p(e_n|e_{n-1}, e_{n-2}) \tag{2.7}$$

Figure (2.1) illustrates the conditional dependency between consecutive observations for both order 1 and order 2 Markov models.

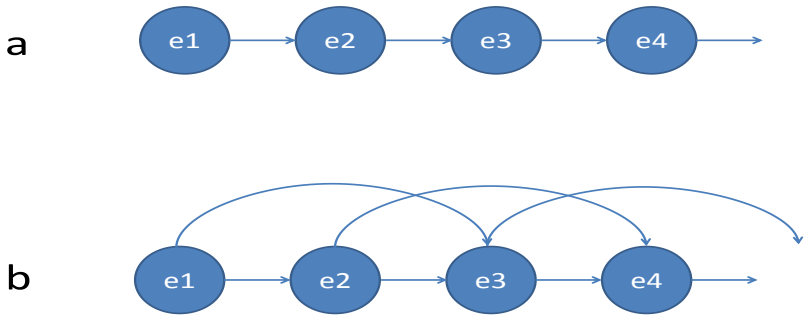

Figure 2.1: Two Markov chains representing the dependency of an observation of n previous observations. **a)** is a first order Markov chain, **b)** is a second order Markov chain (following the style of [Russell and Norvig, 2010, p 568]).

An FOMM of order n can detect patterns of length n. Patterns of higher length $> n$ remain undetected. In order to detect patterns of a length higher than n the order of the underlying FOMM must be increased. The number of parameters in the order n Markov model grows exponentially with the growing order n. If the observations have K different states then the order n Markov model needs $K^{n-1} * (K - 1)$ parameters, which represents a serious limitation for higher order Markov models. Markov models of a higher order require a large amount of training data in order to detect as many patterns as possible. Training a higher order Markov model with insufficient training data causes the model to suffer from both cold-start and zero-frequency problems. Unobserved patterns that first appear during the test phase have both a frequency and probability of zero.

2.2.3.2 DBN Approaches for Prediction

FOMM models treat all the observations to be drawn from the same distribution, but many data sets contain a mixed distribution consisting of n components. Dynamic Bayesian Networks (DBN) [Russell and Norvig, 2010] are a class of predictive mobility models that use two models, namely the state (transition) model and the sensor model (figure 2.2). The state model has an internal (latent) state variable X_t, which is usually unobservable and its state changes at different time steps t. The state model with the conditional distribution $P(X_t|X_{t-1}) \sim \mathcal{N}(AX_t, B)$ captures the system dynamics through defining conditional probabilities on the internal state variable X_t (assuming a prior distribution of the state variable is available), which is restricted by the Markov assumption. Each state defines a distribution from which evidence (output parameters) can be observed. The observed evidence represents observation variables, which together form the sensor model. The sensor model with the condition distribution $P(E_t|X_t) \sim \mathcal{N}(CX_t, D)$ defines conditional probabilities over both internal state variables X_t and the observation variables E_t. The DBN approaches drop the limitation that an observation is dependent only on the most recent n observations, because the observations are connected through the states of the internal state variable, which means a non-blocking path exists between any two observations e_n and e_m via the states of X_t [Bishop, 2007, P 606], thus the current observation depends on the state of X_t (distribution) that caused this observation [Russell and Norvig, 2010, p 568]. Both state and evidence variables are of an arbitrary dimensionality, where the dimension of the latent variable $X_t \in \mathbb{R}^n$ may differ from the dimension of the evidence variable $E_t \in \mathbb{R}^m$.

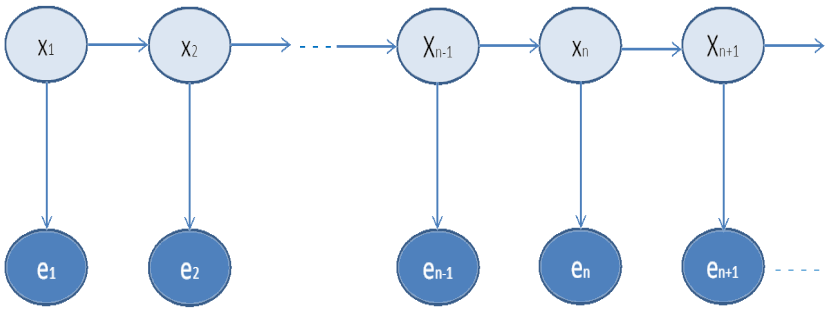

Figure 2.2: A Markov chain containing both observations E and latent states X. Each observation e_i is conditioned on the latent state the observation belongs to. This graphical structure (sometimes called the independence diagram) builds the foundation of Dynamic Bayesian Networks (following the style of [Bishop, 2007, P 606])

The conversion of raw measurements (coming from n sensors) to location measurements plays an important role when using DBN approaches for modeling human mobility [Krumm et al., 2003]. The conversion should be done using a proper deterministic conversion function modeling the relationship between the state variable X_t and evidence vector E_t at time step t [Krumm et al., 2003]. Different models use different conversion functions [Krumm et al., 2003]. All DBN approaches consist of at least one of the following two steps:

- The **inference** step can estimate the state of the model at a given time step

t using a sequence of observations.

- The **correction** step contains an a posteriori estimation of the model parameters upon receiving a new (noisy) observation using the deterministic conversion function.

The inference step of a Dynamic Bayesian Network allows the use of a set of powerful inference types, below we present a short description of four inference types in DBNs (the notion $e_{1:t}$ denotes the observation sequence between time step 1 and time step t as used by [Bishop, 2007]):

- **Filtering -** filtering at any time step t delivers an estimate of the state variable, based on the most recent n states and the observations made up to time t. The estimation is done by computing the conditional probability $p(X_t|e_{1:t})$. Filtering at any time step t allows the current state of the model to be tracked. An efficient procedure for calculation $p(X_t|e_1, e_2, ..., e_t)$ is the forward algorithm.

- **Prediction -** this inference allows the propagation of state estimates into the future, i.e. predicting k future states based on the observations up to time step t through computing the a posteriori distribution of $p(X_{t+k}|e_{1:t})$ for some $k > 0$. The forward algorithm is also an efficient procedure for calculating $p(X_{t+k}, E_t)$.

- **Smoothing -** given the observations up to time step t, smoothing calculates the state of the model at time step k where $k < t$ through estimating the conditional probability $p(X_k|e_{1:t})$. The backward algorithm is an efficient procedure for the smoothing task.

- **Most likely explanation -** this inference type can find the most probable state sequence that might have generated a given sequence of observations, i.e. estimating the conditional probability $p(X_{1:t}|e_{1:t})$. The Viterbi algorithm delivers an efficient procedure for this estimation

Combining both inference and correction steps allows the model parameters of a DBN to be learned iteratively. At each iteration both transitions between states, and the probability that an observation is generated in any state, are estimated. The model parameters are updated based on these estimates, the updated model in turn provides new estimations. These two steps are repeated until convergence is attained. This procedure is an instance of the expectation maximization (EM) algorithm [Russell and Norvig, 2010, P 570-571]

In the following we introduce a few DBN approaches from two categories. A category contains a dynamical linear system having a continuous state variable with Gaussian distribution, whereas the other category contains models that drop the limitation of having a Gaussian distribution.

2.2.3.3 Hidden Markov Models (HMM)

Hidden Markov Models (HMM) drop the Gaussian assumption and are restricted to a single, discrete state variable X and usually a fixed Markov order [Krumm et al., 2003]. Even if more state variables are present, they can be merged to one big state variable with a higher dimensionality. At any time step t, the states of the latent variable of an HMM represent a mixture distribution given by the conditional probability $p(e_t|x_t)$, which has generated the observation e_t. The conditional probability $p(x_t|x_{t-1})$ indicates that the component that generated the observation e_t depends on the component that generated the previous observation e_{t-1} (assuming a first order Markov assumption). The state dynamic can be easily represented by the transition matrix T of size $S \times S$ (S is the number of possible states for the state variable X), where each element T_{ij} of T represents the probability of the transition from state x_i to state x_j. The sensor model can be represented by the diagonal sensor matrix O, because the value of E_t is known at any time step, thus it is only necessary to specify which state has caused the observation E_t to be generated. Once the state and sensor models are specified, the forward α and backward β messages become a simple vector multiplication and can be computed recursively as follows:

$$\alpha_{1:t+1} = aO_{t+1}T^{\top}\alpha_{1:t} \tag{2.8}$$

$$\beta_{k+1:t} = TO_{k+1}\beta_{k+2:t} \tag{2.9}$$

where a is the normalization factor. Both α and β can be combined to compute the transition probability between two states [Russell and Norvig, 2010, P 579] [Duda et al., 2001, P 137-138].

The use of HMM poses different challenges and constraints. The first constraint is the number of the states of the state model, which must be explicitly known, which represents a handicap for a highly dynamic domain such as human outdoor mobility. Another challenge that must be overcome is the dimensionality of the latent state variable, every new dimension leads to an exponentially growing state space. Each new state represents a new distribution over the observations $p(e_t|x_t)$ and the transition matrix must be modified in order to add new transition probabilities to all other states of the state model. An HMM with a large state space requires a large amount of training data to avoid the sparsity and zero-frequency problem and thus the inclusion of further context elements can be very problematic. Furthermore, the fixed order of a standard HMM may not fit well with the complex dependencies in human mobility patterns. Learning the exact parameters of an HMM and the complexity of the solution represents an intractable task for HMM, because none of the known learning algorithms can find the exact solution, but instead a local maxima [Ron and Singer, 1996]. Although HMMs provide flexible structures that can model complex sources of sequential data, dealing with HMMs typically requires considerable understanding of and insight into the problem domain in order to select the initial parameters, so that the model can converge to the globally optimal probability distribution [Begleiter et al., 2004]. Unfortunately it is seldom that an HMM converges to the true probability distribution, therefore the "hidden" part of HMM just complicates the model unnecessarily [Ron and Singer, 1996].

2.2.3.4 Particle Filtering

As exact inference in an unrolled DBN is difficult (see e.g. [Russell and Norvig, 2010, P 590]), methods of approximate inference have to be taken into account. Particle Filtering is a model that allows an approximate inference. Particle Filtering represents the current state at time step t as a set of n state samples associated with a scalar weight, where a higher weight indicates a more probable state according to the transition model [Russell and Norvig, 2010, P 596-597]. Particle Filtering can be augmented with Rao-Blackwellization in order to control the number of samples [Krumm et al., 2003, Gustafsson et al., 2002]. Inference in Particle Filtering works in three steps:

1. Forward propagation by sampling the next state from the N weighted samples based on the transition model.
2. Weighting the new samples by the likelihood of the new observation to be generated by this state.
3. Choosing new N samples from the current state distribution by sampling with replacement, the samples are drawn randomly proportional to their scalar weight.

[Liao et al., 2003] has proposed a mobility model based on Particle Filtering in an indoor environment, but using Particle Filtering for an outdoor environment with a very large state space (over thousands of kilometers) is intuitively intractable.

2.3 Location Acquisition Technologies

Mobile devices (such as mobile phones, PDAs, IPADS and laptops) carried by mobile users may support one or more location acquisition technologies. Generally, depending on the spatial environment, location acquisition technologies can be divided into two groups, namely local (indoor) and global (outdoor) environments.

2.3.1 Local Positioning Technologies

Local positioning technologies have a limited spatial coverage ranging from a few meters to a few hundred meters. They can provide high resolution location data with an accuracy of about 3 meters [Bahl and Padmanabhan., 2000]. Wireless Lan & Bluetooth communications are examples of local positioning technologies. Although both technologies can work everywhere without limitation, their main disadvantage is the expensive infrastructure, which makes them less suitable for use in outdoor environments.

2.3.2 Global Positioning Technologies

Global System for Mobile Communication (GSM)

GSM is the communication standard for mobile phones. Mobile phones communicate with GSM base stations covering an area called a cell. Each cell has its own unique ID. The cells of a GSM net overlap, thus the same physical location is covered by one or more cells. An important advantage of GSM is its presence in buildings and urban canyons. Data collected over GSM suffer from their low resolution, because each cell covers an area ranging between a few hundred meters (Urban) to a few kilometers (rural). Each cell tower covers on average an area of approximately 3x3 km [Eagle et al., 2007, Eagle et al., 2007], which means each cell covers more than one significant location, for example a cell in Seoul is found to contain on average 11 ± 2 locations [Chon et al., 2012].

Global Positioning System (GPS)

GPS increasingly is becoming the standard positioning technology. GPS guarantees to provide location and time information anywhere, at anytime and in all weathers when an unobstructed line of sight is available to at least four satellites. GPS needs a very accurate clock to achieve a perfect estimation of position, because due to the high speed of light, an error of about 10^{-6} seconds corresponds to an error of about 300 meters in position estimation. GPS can estimate the position of a target with a precision of about 10 meters. Differential GPS (DGPS) is an enhancement of GPS which improves the precision of position estimation to about 10 centimeters. GPS unfortunately suffers from signal loss in indoor environments (buildings) and urban canyons [Zheng et al., 2009a].

Other Positioning Technologies

Other less popular system are the Russian Global Navigation Satellite System (GLONASS), the European Union Galileo positioning system, the Chinese Compass navigation system, and the Indian Regional Navigational Satellite System.

2.4 Significant Locations

Global positioning systems can specify any point on the earth using polar coordinates consisting of latitude, longitude and altitude measurements. Prediction of next location using the (raw) polar coordinates is sufficient for many uses such as vehicular tracking, where the next location of a vehicle has to be predicted/calculated within a few seconds. In other use cases where the prediction of location is being made far into the future (hours or even days) such as human outdoor location prediction, it is almost impossible to make an exact inference. Many use cases require discrete representation of locations significant to the user such as "home", "work", "restaurant x", etc. People spend almost 87% of their time at locations and only 13% moving between these locations [Chon et al., 2012]. Thus the locations where

people do spend time are of great importance in understanding their mobility behavior. The main properties that characterize these locations are the amount of time a user stays there and the frequency with which a user visits them. We refer to a location, which is visited by a user with a certain frequency and a certain stay time, with significant location of the user.

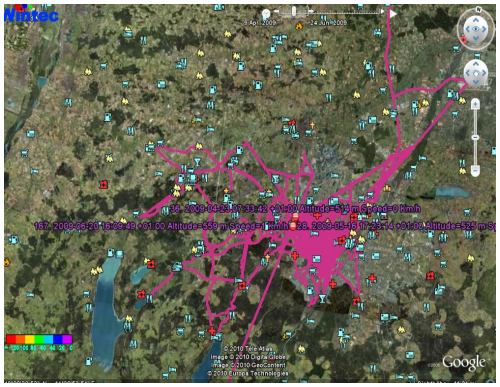

Figure 2.3: The trails of a user in Munich over a period of one year.

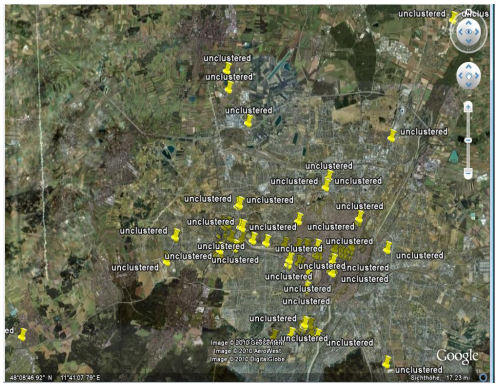

Figure 2.4: The stay points contained in the trails from figure 2.3.

This section provides techniques for transforming raw (continuous) sequences (trails) of GPS data points into sequences (trajectories) of significant locations depending on the duration and frequency of visits. A trail consists of a sequence of consecutive GPS data points of a mobile user. Each GPS data point is a time-stamped record containing at least the current time and the polar coordinates of the user (latitude, longitude and altitude) (figure 2.3).

The time elapsed between two consecutive data points is regarded as the stay time of the user at the first data point. The **stay time** is the amount of time a user spends at a specific location at a stretch before moving to another location. According to the stay time, a trail contains two types of data points, namely stay points with a minimum stay time greater than a threshold and transitory data points for the remaining data points. A **stay point** is a location where the user spends an amount of time which is at least greater than a certain threshold δt (figure 2.4). A **significant location** is a location visited with a frequency of at least $minPt$ where for each visit the user spends at least a time greater than the threshold δt.

GPS measurements are not precise and contain an error up to 10 meters, therefore, the location logged by the mobile device is not necessarily the same, even if the mobile user stays at exactly the same location. Thus a clustering algorithm is necessary for classifying the stay points to a set of the user's significant locations. The transitory data points have a relatively small stay time, therefore they don't represent significant locations. Further, leaving the transitory data points out of consideration reduces both complexity and misclassification of the clustering algorithm.

2.4.1 Significant Location Detection

Clustering algorithms are techniques for partitioning sample data into subsets (clusters). Samples falling into the same subsets exhibit similarities among themselves

compared to elements in other subsets. Clustering algorithms vary according to the sample data available and the selected similarity measurements. The following criteria are important in the choice of the correct clustering algorithm [Duda et al., 2001, P 517-522]

- Number of clusters - whether the number of clusters in the sample data is known a priori
- Size & structure of the cluster - whether the cluster has a certain size or structure
- Similarity measurement - the similarity measurement used to assign a data point to a cluster
- Category labels of the sample data - whether the cluster centers are known

Each clustering algorithm uses a criterion function that has some of the above criteria as input and optimizes one or more of the remaining criteria. In the case of the number of clusters as input, the criterion function is a function that optimizes the centers of the cluster $\mu_1, \mu_2, ..., \mu_i$.

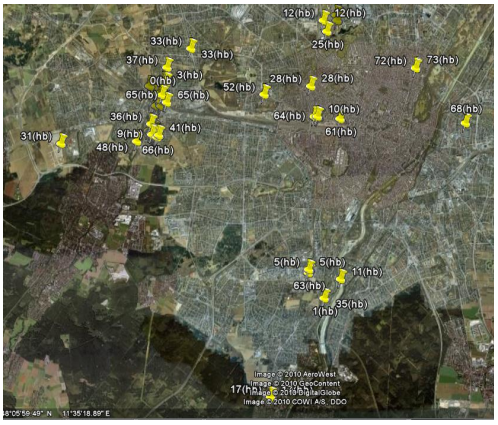

Figure 2.5: The significant locations of the user with a stay time of five minutes (using DB-Scan with $\epsilon = 10$ meters and $minPt = 3$).

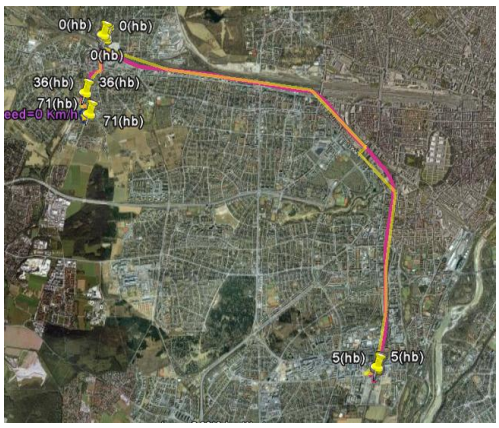

Figure 2.6: The most frequent trajectory of the user (home - bus stop - subway - work - subway - bus stop - home).

Figure 2.5 shows the detected significant locations for a user in Munich using trails over a period of one year and a DB-Scan as a clustering algorithm, figure 2.6 shows the most frequent trajectory of the user.

2.4.1.1 k-Means Clustering

K-means clustering is a procedure for finding a vector $\mu_1, \mu_2, ..., \mu_i$ with k means (centers) representing k clusters in the sample data, where k is known (or guessed) a priori. The criterion function is a function that minimizes the Euclidean distance. If the number of clusters in the sample data is provided, a maximum likelihood procedure can be used to compute the k-means as follows:

Initially k data points are chosen randomly from the data set, these represent the centers of the k clusters. The Euclidean distance to the k centers is calculated for each data point in the data set, the data point is assigned to the cluster with the minimum Euclidean distance. After assigning every data point to a cluster, the centers of the k clusters are calculated and the procedure is repeated until convergence, for example when the cluster centers stop changing.

[Ashbrook and Starner, 2003] proposed a variation on the k-means clustering algorithm that solves the drawback of knowing k a priori. The algorithm uses the cluster size as an input parameter. Instead of k the algorithm uses a radius r and a point from the data set. The clustering procedure is defined as follows:

At each iteration, a data point from the data set is chosen as the center μ_i of the i-th cluster. All the points in the data set that are within the radius r from the μ_i are assigned to that cluster. At the end of the iteration the center μ_i is calculated from the data points assigned to that cluster and the next iteration is started. This process is repeated until μ_i stops changing. μ_i is added to the list of the detected clusters, all the data points within r from μ_i are marked with a label and no longer need be considered. The procedure is repeated until all the data points in the data set are labeled with a cluster label and we are left with a vector of cluster centers $\mu_1, \mu_2, ..., \mu_i$

2.4.1.2 Online Clustering

The previous algorithms assume availability of a full data set prior to the start of the clustering algorithm. In lots of cases comprehensive data is not available, for example in the case of memory lack or when the clusters need to be used while the data streams in. Therefore, algorithms that need the number of clusters to be provided a priori are unsuitable. Furthermore, algorithms that depend on determination of the cluster centers are very sensitive to online changes because even very small changes can cause the whole vector of cluster centers to change. A generic approach called **leader-follower** clustering solves the drawbacks of knowing a priori the number of clusters or the minimum number of data points within a cluster [Duda et al., 2001, P 517-522] as follows:

Given a set of candidate clusters C, a similarity criterion, a learning rate such as the minimum number of data points in a cluster, and a threshold σ, each new data point can be assigned to a candidate $c_i \in C$ if the similarity value of the data point to the candidate $c_i \in C$ is smaller/greater than the threshold σ. The candidate c_i is assumed to be a cluster if its size exceeds the learning rate.

[Kang et al., 2004] proposed the time based clustering algorithm, which is another variant of the k-means algorithm, which can be used online. The algorithm has two parameters, r is radius of the cluster the same as the variant of [Ashbrook and Starner, 2003], the second parameter is the minimum stay time δt at a cluster. The algorithm works as follows.

While a mobile user stays at a significant location, the mobile device logs coordinates near to the center of that significant location. As soon as the user begins to move away from the significant location, a few smaller clusters are formed, but

do not form a significant location, because their stay time is less than δt. A new cluster is formed as soon as enough consecutive data points within a radius r are logged and when the sum of their stay times exceeds $\Sigma t_i > \delta t$. Online Clustering has the disadvantage of not considering the frequency of visit while determining the significant locations of a user.

2.4.1.3 DBSCAN

Density-based Spatial Clustering of Applications with Noise (DBSCAN) was first proposed by [Ester et al., 1996]. It has two input parameters, ϵ is a distance threshold for the maximum distance between two neighboring data points of the same cluster, n is the minimum number of neighboring data points in the vicinity (within a distance threshold ϵ) from a certain data point. The algorithm is based on the notion of density reachability is as follows:

A data point p_i is directly density-reachable from a data point p_j if the following two conditions are true:

- The distance between p_i and p_j must be within the threshold $|p_i - p_j| \leq \epsilon$
- p_i has at least a set N of neighbors containing at least $|N| \geq n$ elements within ϵ distance, i.e. $\forall \acute{p} \in N, |\acute{p} - p_i| \leq \epsilon$

Two data points p_i and p_j in a chain $p_i, p_{i+1}, p_{i+2}, ..., p_{j-2}, p_{j-1}, p_j$ of data points are density-reachable if p_{i+l+1} is directly density-reachable from p_{i+l}. Two data points p_i and p_j are density-connected if there is a third data point p_k such that both p_i and p_j are density-reachable from p_k.

A cluster is a subset of data points C that fulfills two conditions. First each pair of data points $p_i \in C$ and $p_j \in C$ is density connected. Second a data point is a part of the cluster, if it is density-connected to any data point in the cluster. C is called a density-connected set and the data points in C are called core data points. All other non-core data points are considered to be noise, thus DBSCAN can handle noisy data. The output of the algorithm is a set of clusters fulfilling the aforementioned conditions.

2.5 The Choice of a Mobility Model Approach

The quality and performance of a location predictor is highly dependent on the underlying mobility model, hence the choice of a proper approach for the mobility model is a critical issue and must be carefully considered. In addition to performance a set of selection criteria exists, which comes in part from the requirements of the application domain and in part from the properties (strengths and limitations) of the underlying mobility model approach. In addition to performance, these criteria form a framework for the evaluation of the mobility model

2.5.1 Requirements & Scopes

The scope of next location prediction and the requirements, that should be fulfilled by the chosen mobility model in this thesis are presented below

- At any point in time t, the model should predict the **next significant location** of a mobile user, i.e. the location of a user's next stay after leaving their current location. The time of the movement between two significant locations l_i and l_j depends on the stay time of the user at the starting location l_i. The stay time varies between a few minutes to a few hours and in some cases even a day or longer, thus the model should exploit regular patterns over various time-scales in order to predict the next significant location of the user up to one day in the future. Note - both cycle transitions (a user is at a location and remains there) or non-significant locations between start and destination locations of a movement (such as any point in space on the way to a destination location) are not subjects of prediction and thus are left out of consideration.
- **Extensibility -** the model must be **extensible**, to allow for the incorporation of more features such as features from temporal and social contexts, mobile sensor data, calendar etc. Further, the model must be able to capture the conditional interdependency between the included features.
- **Trainability -** as already mentioned in the introduction, the duration of location data storage is a critical issue for privacy. Mobility models that need a large amount of training data reduce the user's acceptance of LBS. Furthermore, such mobility models are hard to retrain if the mobility behavior of a user changes due to a move for example. Finally it must be kept in mind that the availability of sufficient training data is a critical issue and it involves a great deal of expense.
- **Complexity -** a low complexity so that the model can be rendered on mobile devices with possibly limited CPU and storage capacities.

2.5.2 Design and Evaluation Criteria

Memoryless mobility models and all random mobility models are not able to learn the mobility patterns of a mobile user from the user's location history, further they produce movement paths that are not realistic for human mobility behavior (as stated earlier), hence they do not fulfill the above requirements. Mobility models based on exact inference are able to predict the future location of a mobile user for a few seconds or a few minutes into the future. Due to the requirement that the model has to predict the next significant location of the user, which might be hours or sometimes a day in the future, these models are not appropriate candidates. It is almost unfeasible to reliably predict the exact location of a user further into the future for more than few minutes.

Mobility models based on DBN approaches are good candidates because they can handle the uncertainty that arises due to measurement errors, and due to predicting far into the future as well. DBN approaches are able to find patterns of an arbitrary

dimension in the location history of the user. Nevertheless, these models have critical features that have to be well considered. The critical features of DBN approaches expand the evaluation framework for our mobility model.

2.5.2.1 Feature Inclusion

The inclusion of new features is a critical issue of many mobility models based on DBN approaches, because adding new features to the state space causes an exponential increase in the number of states and the size of the transition matrix, which in turn increases the need for more observations and causes the emission matrix to grow accordingly [Krumm et al., 2003]. Adding new features to the observations causes a rapid growth in the size of the observation space, each of which needs an estimation of its probability given the state of the model $p(E|X)$, thus the emission matrix grows exponentially and again the need for more training data increases accordingly.

2.5.2.2 Context Length (Order)

Mobility models based on DBN approaches use conditional probabilities of the form $p(q|s)$ in order to predict the next state q of the model, where $q \in \Sigma$ is chosen from the alphabet of the model Σ (labeled states), $s \in \Sigma^N$ represents the current context and $n = |s|$ is the length of s and is referred to as the order of the model. The order of the model correlates with the number of patterns the model can detect, the higher the order, the higher the amount and the length of possible pattern. The size of the transition matrix of an Nth-order mobility model grows exponentially with the increasing order of the model $|\Sigma|^N \times |\Sigma|$. The use of a higher order has the disadvantage of increasing the need for more training data [Begleiter et al., 2004]. Successful training of mobility models such HMM and FOMM usually requires a very large amount of training data [Abe and Warmuth, 1990] (as cited by [Begleiter et al., 2004]). Insufficient training data causes both the transition and emission matrices to be sparse, thus these mobility models suffer from the so-called cold-start and zero-frequency problems.

2.5.2.3 Zero-Frequency

Cromwell's rule suggests avoiding the assignment of a prior probability of 0 or 1 to a hypothesis (Except when applied to statements that are logically true or false). Otherwise, the influence of any evidence, no matter how strong, on the posterior probability is nullified (by Bayes' theorem, the posterior probability is forced to be 0 or 1). The rule of succession introduced by Laplace suggests defining a mass for the probability of an unseen event in order to be able to adjust the situation once evidences for that event appear [Zabell, 1989]. This problematic is well-known in literature as the zero-frequency problem. Many mobility models based on DBN approaches (including HMM) contravene Cromwell's rule and suffer from the zero-frequency problem. These mobility models need a massive amount of training data to alleviate the impact of the zero-frequency problem [Abe and Warmuth, 1990].

2.5.2.4 Training & Adaptability

The training problem of mobility models based on DBN generally (and HMM specifically) is to find an approximation for an unknown probability distribution. Training mobility models based on DBN approaches is known to be computationally demanding, for example [Abe and Warmuth, 1990] has shown that HMM is not trainable in a polynomial time in the alphabet size unless $RP = NP$. Furthermore, none of the known learning algorithms can find the exact parameters, rather a local maximum solution. Therefore, selection of the initial parameters is of eminent importance in order to find an approximation near the true probability distribution. Selection of the initial parameters typically requires a deep understanding of the problem domain. It is questionable as to whether the performance gained after training the model is significant in comparison with that from the initial parameters chosen. As stated earlier in section (2.2.3.3), it is rare that the learning algorithm converges to the true probability distribution, therefore some of the parameters such as the "hidden" part of HMM just complicate the model unnecessarily [Ron and Singer, 1996].

Human mobility is highly dynamic due to the explorative nature of human beings and due to the continuous changes in their lives, which affect their mobility extremely. Both continuous learning and the adaption of a trained mobility model to new situations should be a priority consideration when designing a mobility model. Adapting many mobility models based on DBN approaches to new situations requires great effort and in many cases even a long and arduous journey of retraining the model, especially when using "hidden" parameters for learning mixed probability distributions.

2.5.2.5 Missing Data Handling

Construction of an optimal solution based on machine learning approaches, can be achieved if the full probability structure of a problem is known [Duda et al., 2001, P 54-56]. Training data may contain errors. The errors have various sources such as sensor errors (non-response or sensor defect), time, human mistakes (leaving some questions unanswered, not recording every movement, shutting down wearable devices or forgetting to charge a device) or errors in interpretation of the data by experts. These sources of errors lead to either complete data records being missing for a period or some features of the data being absent (especially when using multiple sensors). Fixed order mobility models require all features to be present at any time t, therefore they are very sensitive to missing data, especially when a higher order is used. Therefore, the treatment of missing data is of great importance, so that the model can achieve the desired performance.

A well-studied missing data handling method is data imputation [Rubin and Schenker, 1986, Raghunathan et al., 2001, Elliott and Stettler, 2007, Horton and Lipsitz, 2001, Little, 1988, Little and Rubin, 2002, Wu et al., 2004, Timm et al., 2003]. Imputation can be done based on estimating the missing values (based on the probability distribution of the missing attributes), or through adding new weighted examples of a sample that contains missing values. Interpolation (for estimating

complete records which are absent), mean or most frequent values (for estimating the missing fields or the missing complete records) [Little and Rubin, 2002, Little, 1988] are a few methods for data imputation.

Handling missing data can provoke another drawback, namely overfitting, when the model describes the noise instead of describing the underlying relationships. Pruning (for example in decision trees) is an approach to reduce the risk of overfitting. The basic idea behind pruning is to simplify the induced hypothesis. Simplification is done by reducing the number of features to be used in the model. Reducing the number of features may lead to a reduction in the number of patterns the model can detect thus decreasing the model's performance. Therefore many mobility models based on DBN approaches suffer from a dilemma. On the one hand, as many features as possible should be used to detect as many patterns as possible, on the other hand, the use of many features increases both the need for training data and sensitivity to missing data.

2.6 Prediction by Partial Matching (PPM)

PPM belongs to the set of general-purpose prediction algorithms based on Variable Order Markov Models (VOMM) over a finite alphabet Σ [Begleiter et al., 2004]. VOMMs are also referred to as Variable Order Bayesian Networks VOBN or Context-Specific Bayesian Networks (CSBN). In contrast to Bayesian Network (BN) models, which predict the value of a random variable $q_t \in \Sigma$ at any time step t based on conditional probabilities of the form $P(q_t|s = (q_{t-n}, ..., q_{t-1}))$ depending on subsets of random variables $s = (q_{t-n}, ..., q_{t-1})$ of a fixed size $n = |s|$, in VOMM models these subsets may vary based on the specific realization of observed variables in the training data called the context s (hence the term CSBN), where n represents only an upper bound for the size of a context $|s|$. Thus VOMM can exploit patterns of variable sizes [Begleiter et al., 2004].

The use of variable orders reduces the dependency of the model on the size of the training data and the sensitivity of the model to missing data. It has been stated in [Begleiter et al., 2004], that although VOMM models are less expressive than HMMs, some of the VOMM algorithms "enjoy tight theoretical performance guarantees, which in general are not possible in learning using HMMs" [Begleiter et al., 2004]. Further, VOMM models are structurally simpler than HMMs, which makes them amenable for analysis and easier to train.

A key application for VOMM is lossless compression, but it can easily adapted to be used for (location) prediction [Begleiter et al., 2004]. We use an adaptation of Prediction by Partial Matching (PPM). PPM is an instance of general VOMM that outperformed other instances in a comparative study by [Begleiter et al., 2004]. VOMM approaches generally use a tree structure to alleviate the problem of transition matrix sparseness. Let N be the order of the model, then the tree has a maximal depth of $N+1$ and each path in it defines a subsequence of symbols appearing in the training data. Each node of the tree is labeled with a symbol q from the alphabet Σ and has a counter c for bookkeeping the number of occurrences of the context constructed through concatenating all the symbols from the root of the tree to that

specific node.

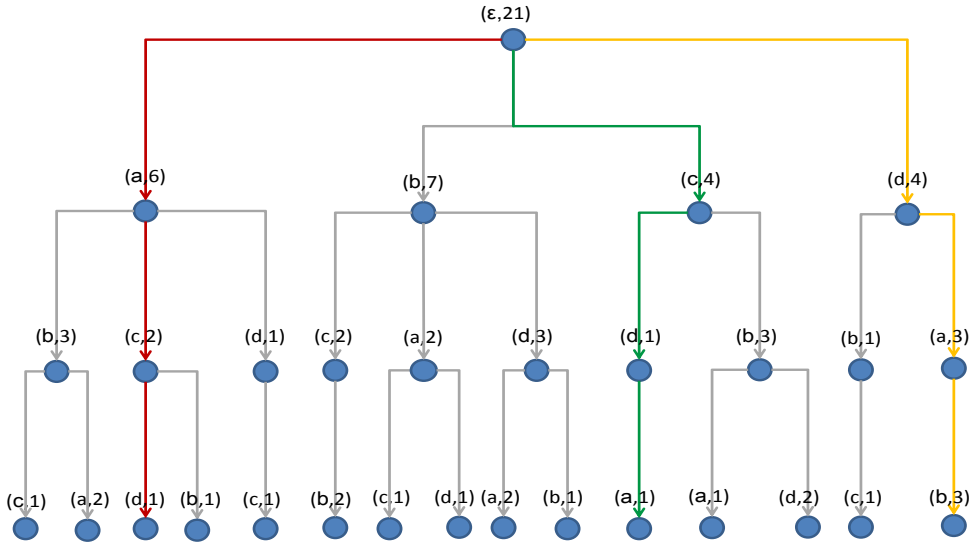

Figure 2.7: The prefix-tree of a PPM VOMM tree using the training sequence "acdabcbdbcbdabacbdabadc" from the location history of a user. The alphabet of the model is the set of symbols occurring in the training sequence, namely ($\Sigma = \{a, b, c, d\}$). The colored paths in the tree represent the first three maximum sub-patterns of the training sequence, namely acd, cda and dab. The sub-pattern dab occurs three times in the training sequence, thus the counter of the corresponding node is set two 3.

Figure (2.7) shows an example PPM VOMM tree (also called prefix-tree), which is constructed by training an order two PPM VOMM model with the training sequence "acdabcbdbcbdabacbdabadc" from the location history of a user. The alphabet of the model is the set of symbols occurring in the training sequence, namely ($\Sigma = \{a, b, c, d\}$). Each node of the tree is labeled with a symbol from the alphabet and the corresponding counter for bookkeeping its occurrence. Each node represents a subpattern constructed by traversing the tree from the root to that node. The nodes on level one represent sub-patterns of length one, the leafs represent maximum sub-patterns of the model of length $N + 1$, where N is the order of the model. The first three maximum sub-patterns of the training sequence are colored, namely acd, cda and dab. The subpattern dab occurs three times in the training sequence, thus the counter of the corresponding node is set two 3.

2.6.1 Training PPM VOMM

The context s_t of the original PPM VOMM described in [Begleiter et al., 2004] has a length of N, where N is the order of the underlying Markov model. At each time step during training phase, we insert a path of length $N + 1$ into the tree according to the current context s_t and the symbol that appears after this context. During the following time steps a new context evolves by appending the symbol

that appeared after the previous context s_t to its end and removing the first symbol from s_t. According to this process, the contexts at two consecutive time steps have $n-1$ symbols in common [Begleiter et al., 2004]. This process is repeated until the end of the sequence.

According to the above process, training the PPM VOMM tree effectively corresponds to instantiating or updating node counters in the tree. Assuming an order of $n = 3$ and a subsequence $ABCD$ from the training data, all the counters along the path from the root ϵ to the node labeled with last symbol D have to be incremented ($s = ABC$).

2.6.2 Prediction Using PPM VOMM

Prediction using PPM VOMM estimates the conditional probability $p(q|s)$ for a symbol from the alphabet $q \in \Sigma$ appearing after the context s of length $|s| = N$. PPM VOMM predicts the symbol q that maximizes the probability $\text{argmax}_q p(q|s)$ as follows. At each time step, the algorithm starts traversing the tree looking for a path matching the current context s_{t_i} having the length $|s_{t_i}| = n$. If no matching path can be found then the algorithm escapes iteratively to find a path matching a maximum sub-pattern s'_{t_i} of the current context s_{t_i} by removing the first symbol, that means $s_{t_i} = bs'_{t_i}$ where b is any symbole from the alphabet $b \in \Sigma$.

To alleviate the drawbacks of the zero-frequency problem, PPM defines an escape mechanism as in [Begleiter et al., 2004]. For all symbols that have not yet appeared after context s, the escape mechanism assigns a probability mass $P(escape|s)$. The remaining mass $1 - P(escape|s)$ is distributed between the symbols appearing after context s. Equation (2.10) determines the probability of any symbol q occurring after context s recursively.

$$P(q|s) = \begin{cases} \tilde{P}(q|s) & \text{if } q \in \Sigma_s \\ \tilde{P}(escape|s)\ P(q|\text{suf}(s)) & \text{else} \end{cases} \tag{2.10}$$

where Σ_s is the set of symbols appearing after context s, $\text{suf}(s)$ denotes the longest suffix of s. The probability of any symbol appearing after an empty context $|s| = 0$ is $P(q|\varepsilon) = \frac{1}{|\Sigma|}$. Hence, PPM VOMM is able to assign a probability mass to any symbol independently of its occurrence in the training sequence. For symbol q and context s, let $C(sq)$ be the counter that counts the occurrences of sq, equations (2.11) and 2.12 are estimations of both probabilities $\tilde{P}(q|s)$ and $\tilde{P}(escape|s)$ respectively:

$$\tilde{P}(q|s) = \frac{C(sq)}{|\Sigma_s| + \Sigma_{q' \in \Sigma_s} C(sq')} \tag{2.11}$$

$$\tilde{P}(escape|s) = 1 - \sum_{q \in \Sigma_s} \tilde{P}(q|s) = \frac{|\Sigma_s|}{|\Sigma_s| + \Sigma_{q' \in \Sigma_s} C(sq')} \tag{2.12}$$

The escape mechanism is a technique to deal with the zero-frequency problem. The summand $|\Sigma_s|$ in the denominator of eq. (2.11) and (2.12) is a Laplace estimator-like mass for assigning a probability mass to observation sequences that does not appear in the training sequence [Bapierre et al., 2011]. Thus PPM VOMM does not

contravene Cromwell's rule (compare with the rule of succession in general statistics [Zabell, 1989]) and can deal with symbols $q \notin \Sigma_s$, where Σ_s is the set of all symbols appearing after context s. More clearly, Σ_s is the set of already known symbols that have occurred after the context s, and it does not contain the new emerging symbol q.

A further important advantage of PPM VOMM is that training and prediction do not have to be separated, the simple mechanism of instantiating and updating PPM VOMM allows the model to be trained and tested simultaneously, which indeed significantly helps to alleviate the drawback of the cold-start problem, as well as to increase the adaptability of the model when the mobility behavior of the user changes.

2.7 Empirical Results

We evaluated the PPM VOMM model proposed in the previous section using two different datasets, namely the GeoLife [Zheng et al., 2009b, Geo, 2013a] and the Reality Mining [Eagle and Pentland, 2006] datasets.

2.7.1 GeoLife Dataset

The GeoLife dataset contains the location histories of 153 users over a period totaling 2 years between April 2007 and May 2010 (the dataset has been updated recently and now contains 183 users over a period totaling 3 years between April 2007 and August 2012 [Geo, 2013a, Geo, 2013b]). The dataset is very dense because for almost 91% of the data points, a pair of consecutive data points are either 1-5 seconds or 5-10 meters of each other [Geo, 2013b]. Each data point is a time-stamped GPS record containing at least the absolute GPS coordinates (latitude, longitude). The dataset contains 69,679 trails. A trail is a sequence of consecutive data points without interruption. On average, the dataset contains $\approx$ 150,000 GPS records per user (table (2.1)).

Type	**All**
# Data Points:	22 825 083
# Users:	153
# Trails:	69 679
Av. Data Points per User:	149 184

Table 2.1: Data points and user statistics.

The data points were collected using different GPS loggers with different data formats, thus pre-processing is needed to convert the different formats to one unified format. In a second pre-processing step we eliminate the transitory GPS-records between two stay points. A stay point is a location where the user spends a period of time greater than a threshold δt as stated earlier.

2.7.1.1 Stay Point Detection

Prior to stay point detection, we group consecutive GPS records of the same trail that occur within a radius smaller than a threshold (δd) to one single data point p. Let $SP = \{p_n, p_{n+1}, ..., p_{n+i}\}$ be the set of consecutive GPS records of the same trail that occurs within δd. The time the user spends at SP is the time elapsed between p_n and p_{n+i}. The coordinates of SP are set to the average coordinates of the data points. The dataset contains $11,254,785$ data points when setting $\delta d = 10$ meters. SP is a stay point only if the time elapsed between p_n and p_{n+i} is greater than δt. Note: GPS capable devices suffer from signal loss inside building and in urban canyons, thus the time elapsed between two consecutive data points p_i and p_{i+1} is considered to be the stay time of data point p_i. The stay points yield locations where the user has spent a period of time greater than δt.

Δt	0	10	20	30
# Data Points:	11 254 785	48 985	38 125	32 586
# Users:	153	127	121	116
Stay Time Rate:	100%	98.7%	98.0%	97.0%
# Trails:	69 679	44 896	36 254	31 496
Av. Stay Points per User:	73 561	323	254	217

Table 2.2: User and stay point statistics using $\delta d = 10$ meters.

On average a user spends most of their time at a few locations, almost 99.0% at locations with a minimum stay time of five minutes $\delta t = 5$ as to be seen in table (2.2). The number of stay points depends on the choice of δt, for example, for $\delta t = 10$ the number of stay points is $48,985$, which corresponds to 0.44% of the total number of data points (table (2.2)), thus detecting the user's significant locations based on stay time leads to a huge reduction in the amount of data to be processed.

According to our definitions in section (2.4), the stay time as a single classification feature is insufficient for detecting all the significant locations of a user. A further classification is needed for detecting the significant locations of the user.

2.7.1.2 Significant Location Detection

Stay points need to be further classified according to the frequency of visits to the same location in order to reduce the amount of false positive and false negative classifications. The stay points need further comparison to classify different stay points to the same physical location. As stated earlier, GPS is known to have an error of $\leq$ 10 meters, even if the user stays at exactly the same location. Thus stay points within a distance smaller than a threshold ϵ are assumed to belong to the same single location (note - stay points need not be consecutive, they need not even to belong to the same trail). Setting ϵ to a high value results in merging many neighboring locations to one single huge location, thus the detected locations are coarse grained compared to setting ϵ to a smaller value. Setting ϵ to a low value causes one and the same location to be detected multiple times as different locations.

Significant locations are visited frequently by a user, the frequency of visit n varies according to the type of location. A user visits their home and work locations very frequently, whereas they visit their favorite bar, restaurant, boutique or a swimming pool in their neighborhood with a lower frequency. The dental practice or a favorite Christmas market with an even lower frequency. Setting the frequency to a high value has the consequence of not detecting significant locations visited with a low frequency, but on the other hand setting the frequency to a low value leads to high false positive classifications, because an outlier like a traffic light at an intersection may be detected as a significant location. We aim to detect as many significant and fine grained locations as possible even if they are visited with a low frequency. On the other hand the detected locations should contain as few outliers as possible.

Density based clustering DBSCAN as introduced previously in section (2.4.1.3) can detect locations of an arbitrary shape and has the ability to detect outliers, therefore we prefer DBSCAN for detecting significant locations. The parameter $minPt$ is the parameter that indicates the minimum number of neighbors a stay point must have within a radius smaller than ϵ inside a cluster in order to be assigned to this cluster. Otherwise the stay point belongs to another cluster or it is considered to be an outlier. We fix the value of the parameter $minPt = 2$, i.e. a stay point SP_i belongs to a significant location l, if SP_i has at least two neighbors inside l. Table (2.3) shows the number of detected significant locations with different parameter configurations for ϵ, δt and the minimum frequency with which a user visits a location l, so that l can be considered a significant location. The number in braces represents the average frequency of visit for the detected significant locations.

Min. Stay Time (δt)	ϵ \ Frequency (minPt)	2	5	10
10 Minutes	10	3 535 (6.0)	485 (29.4)	231 (55.0)
	30	3 888 (7.6)	784 (28.5)	362 (54.2)
	50	3 997 (8.2)	900 (28.2)	425 (52.6)
20 Minutes	10	2 660 (6.3)	372 (30.5)	186 (54.7)
	30	2 821 (8.1)	615 (29.0)	295 (53.4)
	50	2 904 (8.7)	700 (28.8)	340 (52.3)
30 Minutes	10	2 247 (6.4)	328 (30.4)	157 (56.6)
	30	2 331 (8.5)	542 (28.8)	263 (52.5)
	50	2 386 (9.1)	605 (29.1)	303 (51.5)

Table 2.3: The Results DBScan started with different parameter settings (ϵ, $minPt$ and δt).

The amount of time spent at the detected significant locations varies according to the parameter settings, a user spends on average 82.85% of their time at locations identified with the parameter setting ($\delta t = 30$, $\epsilon = 10$, $minPt = 2$) and 87.95% with the parameter setting ($\delta t = 10$, $\epsilon = 50$, $minPt = 2$). On average for all the parameter settings, the users spend $85.6 \pm 1.7\%$ of their time at the detected significant locations, which is entirely in accordance with the findings in [Chon et al., 2012]. This means the mobility of a user is predictable for almost 85.6% of the time if the mobility model is able to predict their significant locations.

2.7.1.3 Prediction Accuracy

The number and the frequency of visits to the significant locations varies according to the parameter settings, hence the prediction accuracy varies accordingly. In order to reduce the effect of significant locations with high stay times such as home and work locations, we ignore cyclic transitions for the calculation of prediction accuracy. Each correct prediction is weighted equally, no matter how long the stay time is.

Min. Stay Time (δt)	Frequency (minPt) / ϵ	2	5	10
10 Minutes	10	50.4	61.6	67.8
	30	49.3	59.5	65.0
	50	49.1	59.5	64.6
20 Minutes	10	52.6	64.3	70.0
	30	53.1	62.7	67.9
	50	52.7	61.9	68.6
30 Minutes	10	55.2	68.7	75.4
	30	54.9	64.3	71.3
	50	54.5	64.2	71.2

Table 2.4: Prediction accuracy in percent for the GeoLife Dataset.

Table (2.4) shows the prediction accuracy for different parameter settings of the clustering algorithm. The prediction accuracy varies between 49.1% and 75.4%.

Google maps provides functions for finding the nearest natural address to specified geographical coordinates (Longitude/Latitude) [goo, 2013]. We applied our mobility model to the sequence of natural addresses visited by each user. An average absolute improvement in accuracy of 0.03 could be achieved by making use of Google maps' geo-coding service. The improvement can be attributed to the fact that some addresses cover a large area, such as building complexes of a university campus. In fact, we found that multiple clusters were assigned to the same natural address. This result demonstrates that the clusters have a finer granularity than natural addresses. Therefore we decided to use the clusters for subsequent evaluations.

For further analysis we fixed the parameters to ($\delta t = 10, minPt = 5, \epsilon = 10$). Table (2.5) shows the statistics of the significant locations detected using this parameter setting.

2.7.2 Reality Mining Dataset

The Reality Mining dataset [Eagle and Pentland, 2006] was collected from 100 users over the course of nine months. The dataset contains location information in form of GSM cell tower IDs. Each cell towers covers on average an area (cell) of 3×3 km [Eagle et al., 2007, Eagle et al., 2007]. In order to guarantee total coverage, the same area is covered by multiple cell towers, thus a mobile phone can be observed by different cell towers even if the user stays at the same location. The dataset contains a user specific label for the cell towers that contain significant locations of the users.

	$\delta t = 10, minPt = 5, \epsilon = 10$
# Users:	127
# Significant Locations:	3,535
# Stay Points:	21,292
Av. Visits per Location:	6.02 ± 6.49
Av. Visits per User:	167.65 ± 194
Av. Locations per User:	27.83 ± 27.6
Av. User Entropy	1.57 ± 0.43

Table 2.5: Visits and location statistics with parameter setting $\delta t = 10, minPt = 5, \epsilon = 10$.

The use of user-specific labeling for the cell towers is advantageous because it allows supervised learning, thus errors that occur due to clustering are excluded. Table (2.6) shows some of the statistics for the GSM records for all and for labeled towers.

	All cells	**Labeled cells**
# Users:	90	77
# Cells:	32 656	1 632
# Labels:	-	958
Stay Time Rate:	100%	75.77%
# Data Points:	2 536 034	1 105 004
Av. Data Points per Cell/Label:	77.66 ± 131.8	1 415 ± 2 212
Av. Data Points per User:	28 495 ± 14 234	14 350 ± 8 633
Av. Label per User:	-	12.44 ± 8.44
Av. User Entropy	-	1.48 ± 0.43

Table 2.6: Observations and location statistics of the Reality Mining dataset.

A user in the Reality Mining dataset spends on average 75.77% of their time at locations labeled by them, which is around 10% less than the time spent by an average user in the GeoLife dataset. The reason for the reduced stay times at significant locations is the supervised detection of significant locations. Long stays at non-significant locations such as waiting in traffic jams are excluded by the users themselves, whereas such stay times may be included in the GeoLife dataset, for example when the users are repeatedly stuck in traffic jams at the same location on their usual route to work/home.

The locations in the reality mining dataset unfortunately are very coarse because each cell covers a wide area, the user is expected to be at a location when they are observed by a labeled cell tower, even if they are just in the vicinity of that location. In addition, a cell tower can cover multiple locations, but has only one label, thus the user can be mistakenly expected at a location.

2.7.3 Performance Analysis

[Mathew et al., 2012] has proposed a mobility model based on a hidden Markov model, the mobility model was evaluated using the GeoLife dataset. The performance of their mobility model was 13.85% for the correct location and 26.4% when

the correct location was in the top five most probable locations. The performance of hidden Markov models has also been addressed by [Asahara et al., 2011] and found to be poor for location prediction.

[Eagle and Pentland, 2009] has proposed a work based on principal component analysis (PCA), the performance of the model is evaluated using the Reality Mining data set. Given the behavior of user in first 12 hours of a day, the model is able to predict the behavior of the user in the remaining 12 hours of the day with an accuracy of 79% [Eagle and Pentland, 2009].

[Ashbrook and Starner, 2003] has proposed a mobility model based on a naive Markov model. Unfortunately the authors have not reported any evaluation results, but a similar mobility model has been proposed by [Yu et al., 2006], the mobility model is based on a hybrid Markov model and achieves a performance of $\leq 75\%$. The authors acknowledge that order n Markov model achieves a good performance, this motivated us to implement the order n naive Markov model ourselves in order to compare the performance of our PPM VOMM to. We use the performance of the aforementioned mobility models as bench marks for the performance of the PPM VOMM model.

2.7.3.1 Size of Training Data

The availability of sufficient training data is a critical issue of predictive models based on probabilistic reasoning. The acquisition of training data itself is critical because it is associated with considerable effort and cost. Further, a mobility model with strong dependency on the size of training data is less adaptable, hence less applicable to domains with rapidly changing dynamics such as human mobility. The danger is that the dynamics of the domain change faster than the model is trained, which in turn leads to a poor prediction accuracy of the model.

Dataset	Training Data % / Mobility Model	0.9	0.75	0.5	0.25	0.0
GeoLife	FOMM	52.01	48.29	43.72	39.13	0.0
	PPM VOMM	61.63	61.57	61.51	61.56	60.60
Reality Mining	FOMM	81.58	80.09	78.54	77.39	0.0
	PPM VOMM	81.02	80.80	79.82	79.75	75.90

Table 2.7: Prediction accuracy in percent vs. size of training data, the order in all cases was set to two.

In order to investigate the dependency of both FOMM and PPM VOMM on the size of the training data, we varied the training data during multiple test trials on both GeoLife and the Reality Mining datasets. The ability of PPM VOMM to be trained and tested simultaneously is a key property that makes PPM VOMM less dependent on the size of the training data. Table (2.7) shows that the accuracy of PPM VOMM drops off by 1.03% and 5.13% for both Geolife and the Reality Mining datasets respectively when the amount of training data is reduced from 90% (10% test data) to 0% (100% test data). Reducing the amount of training data from 90%

to 10% drops the accuracy marginally on both datasets, therefore we can conclude that PPM VOMM is indeed substantially less dependent on the size of training data compared to FOMM. FOMM cannot be trained and tested simultaneously. When the amount of training data is reduced from 90% to 10%, the accuracy of FOMM drops off considerably from 52.01% to 39.13% on the GeoLife dataset, which means a reduction of about 12.88%. FOMM loses on average 8.03% of its accuracy on the Reality Mining dataset when the size of training data is reduced from 90% to 10%.

2.7.3.2 Context Length (Order)

A critical issue when using mobility models based on probabilistic reasoning is the context length of the model (order). The length of the context is an indication of the degree of memory of the model, it points to the number of previous states/observations that influence the current state/observation of the model. Table (2.8) shows the prediction accuracy of FOMM and PPM VOMM for both Reality Mining and GeoLife datasets for varying model orders. The prediction accuracies for the GeoLife dataset illustrate clearly the effect of the growing order, the PPM VOMM is significantly less sensitive compared to FOMM because of the variable order of the model. A symbol, that has not yet observed in the training data causes n consecutive false predictions in the FOMM mode. Whereas due to the variable order of PPM VOMM, possible false predictions due to unseen symbols are bound by [1, n]. Whenever a context s of length n is not detected in the training data, PPM VOMM escapes to the next lower order by removing the first symbol of the context s. Two-sided unpaired T-tests for the varying orders confirm the significance of the improvements using PPM VOMM compared to FOMM (T-test values $0.0014, 8.1 * 10^{-7}, 3.5 * 10^{-5}, 4.6 * 10^{-9}$ for order 2, 3, 4 and 5 respectively).

Dataset	Order / Mobility Model	2	3	4	5
GeoLife	FOMM	43.67	33.48	25.95	19.84
	PPM VOMM	61.51	61.85	60.54	59.98
Reality Mining	FOMM	78.54	78.13	76.68	74.17
	PPM VOMM	78.97	78.34	77.2	75.0

Table 2.8: Prediction accuracy in percent of both FOMM and PPM VOMM using different order, the training size was set to 0.5

The prediction accuracies for the Reality Mining dataset are stable for both FOMM and PPM VOMM compared to the GeoLife dataset for the following reasons. The significant locations of the users in GeoLife are biased due to unsupervised clustering of the raw GPS measurements. The significant locations in the Reality Mining dataset, however, are unbiased because the users have labeled them themselves. Further, the locations in the GeoLife dataset are finer grained than the GSM cells in the Reality Mining dataset. It is more difficult to predict finer grained locations than cells of size 3×3 km. That is also why the average prediction accuracy for the Reality Mining dataset is considerably higher than the prediction accuracy for the GeoLife dataset. Generally, the accuracy of PPM VOMM is higher compared

to FOMM, the relative improvement in accuracy on the Reality mining dataset is $\simeq 1\%$ and varies on the GeoLife dataset between 41% and 200%.

In order to be sure that the improvements are not due to artifacts of the GeoLife dataset, we applied PPM VOMM to the MDC dataset [Laurila et al., 2012]. The MDC dataset is also GPS-based and the locations are finer grained and more precise than GSM cells [Laurila et al., 2012]. The raw GPS data is transformed into symbolic locations, each symbolic location covers a radius of 100 meters. The sequence of significant locations contains every symbolic location where the user has spent at least 20 minutes. FOMM achieved an accuracy of 42.72%, whereas PMM VOMM achieved 52.72%, which is about a degree of magnitude better than FOMM, the relative improvement in accuracy is even 23.4%. The performance of PPM VOMM on the MDC dataset is considerably less than its performance on the GeoLife dataset, because the MDC dataset does not consider the frequency of visit while detecting significant locations, thus the average number of locations per user is significantly higher 66.30 ± 33.90 compared to the GeoLife dataset 27.83 ± 27.60. We use the MDC dataset in another context in chapter 6. Therefore we continue our evaluations using the GeoLife dataset in this chapter.

2.7.3.3 Zero-Frequency

The number of patterns detected in a dataset depends on both the size of training data and the order of the underlying mobility model. Observations made for the first time during the test phase are problematic and lead to poor performance of the mobility model. Zero-frequency is a well-known problem of models based on probabilistic reasoning that is closely related to new observations that are not yet seen in the location history of the user. An unseen observation, for example, occurs when a user visits a location for the first time, which can always happen due to the explorative nature of human beings, therefore the consequent problems are inevitable. Although zero-frequency is unavoidable, its negative side-effects can be alleviated by the choice of the proper mobility model.

The negative side-effects of visiting a new location are exacerbated by the order n of the underlying mobility model. The new location is part of n different consecutive contexts s, each of which has a zero-frequency, therefore the negative side-effects are increased multiplicatively by the order of the mobility model. As stated earlier the PPM VOMM uses an escape mechanism during the test phase to alleviate the effects of zero-frequency. Although the PMM VOMM model (at least until now), similarly to HMM and FOMM, is unable to predict the new location, it does have two advantages compared to HMM or FOMM. The escape mechanism of PPM VOMM allows assignation of a probability mass to unseen contexts and its variable order allows predictions using a lower order when prediction is not possible using the higher order. Therefore PPM VOMM is able to predict parts of the following $n - 1$ contexts.

Setting the order to two and the amount of training data to 0.5, 23.73% of the contexts in the GeoLife data set occur only during the test phase, of which 48.84% could be correctly predicted. That corresponds to a 0.1159 overall absolute improvement in accuracy compared to FOMM ($\simeq 65\%$ of the total improvement in accuracy

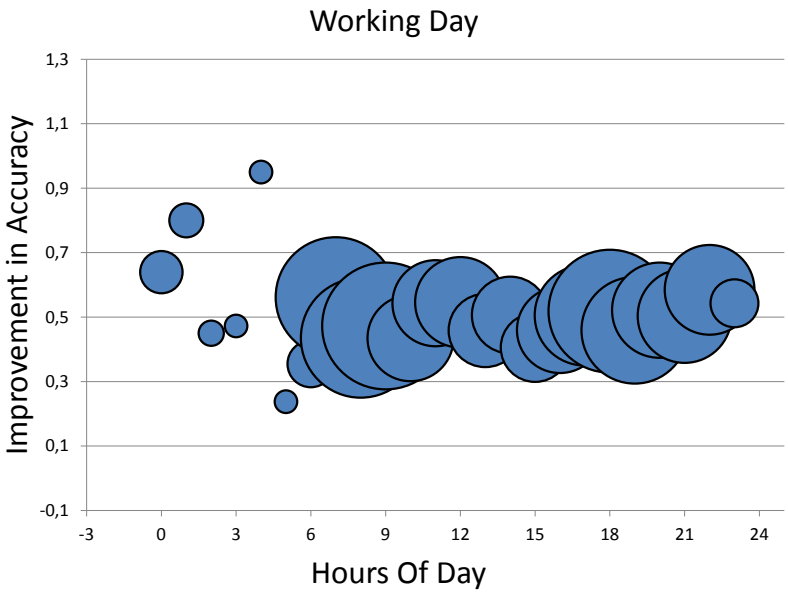

Figure 2.8: The distribution of absolute improvements in accuracy (y-axis) over the hours of work days (x-axis) due to alleviating the negative impact of zero-frequency. The size of bubbles indicate the amount of unseen contexts.

of 0.1784). This result impressively demonstrates the advantages of both variable order and the escape mechanism of PPM VOMM.

The amount of unseen contexts during the hours of a work day are plotted in figure (2.8). Most of the unseen contexts (by looking at the size of the bubbles) occur during two time intervals. The first interval is in the morning hours between 7 and 9 a.m. when people are usually on their way to their work places. Users seem to visit new locations during these hours. An explanation for this could be short stops for coffee/snacks at coffee shops, keeping doctor's appointments or visiting the authorities. The second interval is around the evening hours when people usually leave their work places and possibly explore new locations like a shopping mall, a bar or restaurant before going home.

Unlike work days, the distribution of known contexts over the hours of weekend days (by looking at the size of the bubbles) exhibits a smooth increase that reaches its peak between 9 and 11 p.m. as can be seen in figure (2.9). Users appear to be most explorative during the afternoon and evening hours at the weekend. Indeed this behavior can be generalized to apply to most people, not only the users in this dataset. Because people conduct their free-time activities during the afternoon hours of the weekend and then let the day end comfortably in a restaurant or a bar.

Accordingly, alleviation of the negative impact of zero-frequency improves the accuracy of PPM VOMM during the hours of the day when people are most explorative. Both figures (2.8) and (2.9) demonstrate the improvement in accuracy over the hours of both work and weekend days. At this point we come to the conclusion that PPM VOMM has significant advantages over HMM and FOMM which result

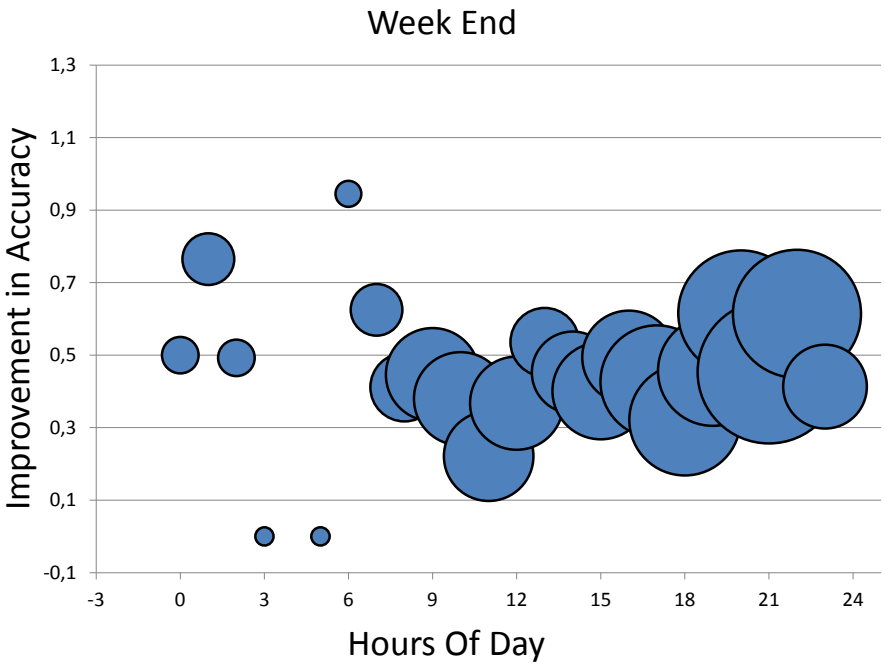

Figure 2.9: The distribution of absolute improvements in accuracy (y-axis) over the hours of weekend days (x-axis) due to alleviating the negative impacts of zero-frequency. The size of bubbles indicates the amount of unseen contexts.

from the reduction in the impact of zero-frequency. Therefore, PPM VOMM can possibly offset the advantages of the greater expressiveness of HMM, albeit considerably simpler in structure and easier to train.

2.7.3.4 Number of Locations

The amount of pattern that can be found in the location history of a user depends heavily on the number of locations that user visits. The higher the number of locations, the higher the amount of pattern. Therefore there is an interdependency between the number of locations visited by the user and the mobility predictability of that user. The mobility of users with a regular life, following a few regular patterns containing a few locations is fairly predictable. The mobility of users who are rather explorative in nature and love to experience new locations that are as yet unvisited by them is less predictable. Hence the dependency of a mobility model on the number of locations is an important evaluation criterion. The aim is to increase the mobility predictability of users who visit lots of locations.

The prediction accuracies of both FOMM and PPM VOMM show strong negative correlations with the number of locations visited by the users. Figure (2.10) illustrates these negative correlations. Pearson's correlation coefficient $r = -0.43$ shows a strong negative correlation, Spearman's rank correlation coefficient $\rho = -0.59$ shows an even stronger negative correlation with probability of error $P(\epsilon) \leq 0.0$.

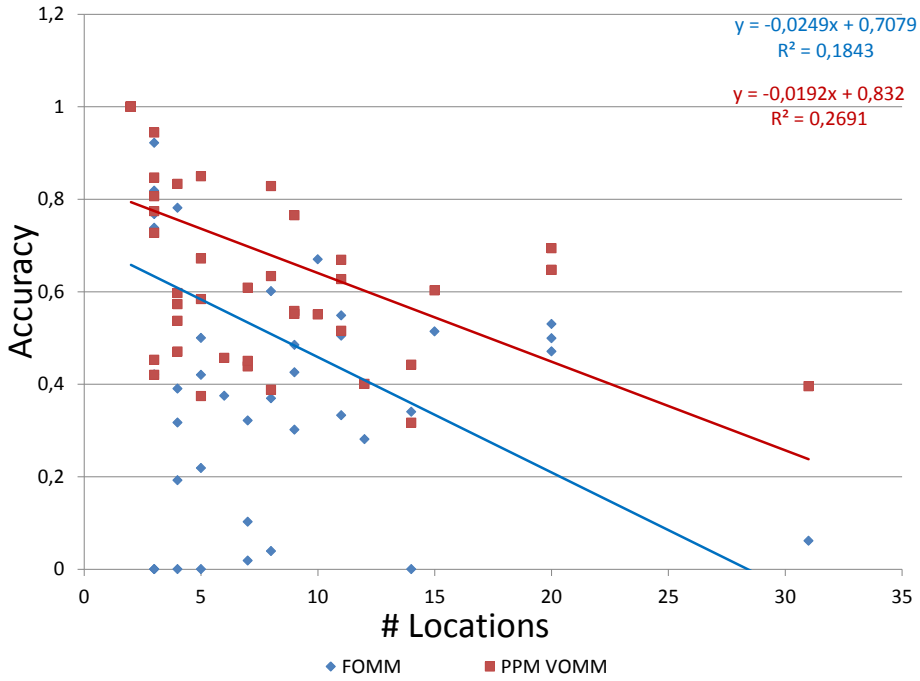

Figure 2.10: The number of locations visited by the users shows a strong negative correlation with the accuracy of both FOMM ($r = -0.43, \rho = -0.59, P(\epsilon) = 0.0$) and PPM VOMM ($r = -0.52, \rho = -0.69, P(\epsilon) = 0.0$).

Although there is a strong negative correlation between accuracy and the number of visited locations, the improvement in accuracy gained by using PPM VOMM instead of FOMM shows a positive correlation according to Pearson's correlation coefficient $r = 0.19$, and also a strong positive correlation according to Spearmann's rank correlation coefficient $\rho = 0.44$ (figure (2.11)). PPM VOMM outperforms FOMM on all scales, but markedly for explorative users with a high number of visited locations. This result implies that PPM VOMM is less dependent on the number of locations compared to FOMM. A two-sided unpaired t-test $= 0.003$ confirms the significance of the improvement in accuracy.

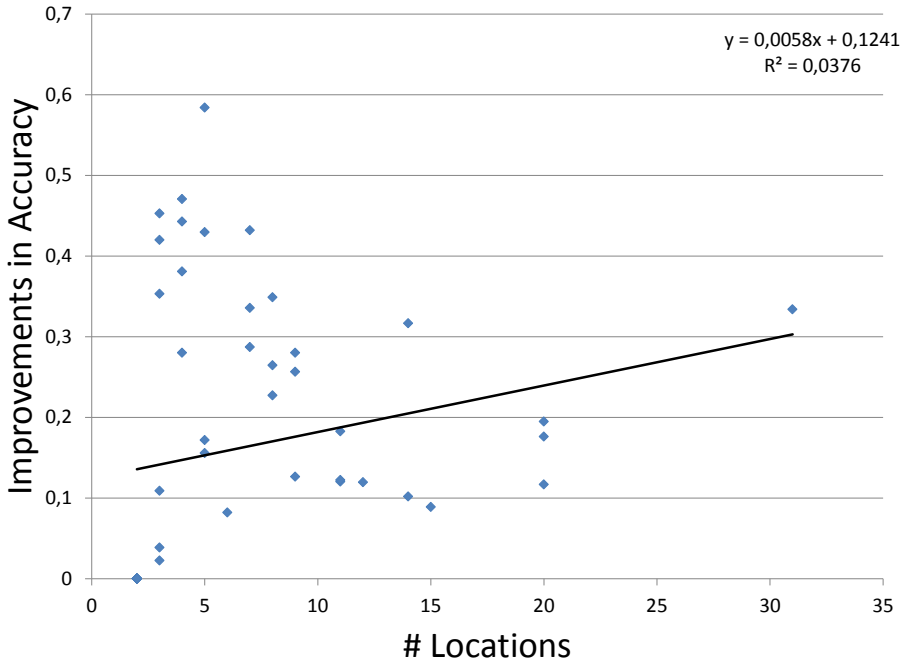

Figure 2.11: There is a positive correlation between the number of locations and absolute improvement in accuracy ($r = 0.19, \rho = 0.44, P(\epsilon) = 0.0014$).

2.7.3.5 Frequency of Visit per Location

The frequency of visit per location is a further measurement that influences the predictability of locations and thus the performance of a mobility model. Locations visited with a high frequency are more predictable than locations visited with less frequency. Most mobility models are able to predict locations with high frequency of visit such as home and work places, however a true indicator of the quality of a mobility model is its ability to predict locations with a low frequency of visit such as a favorite restaurant of the user.

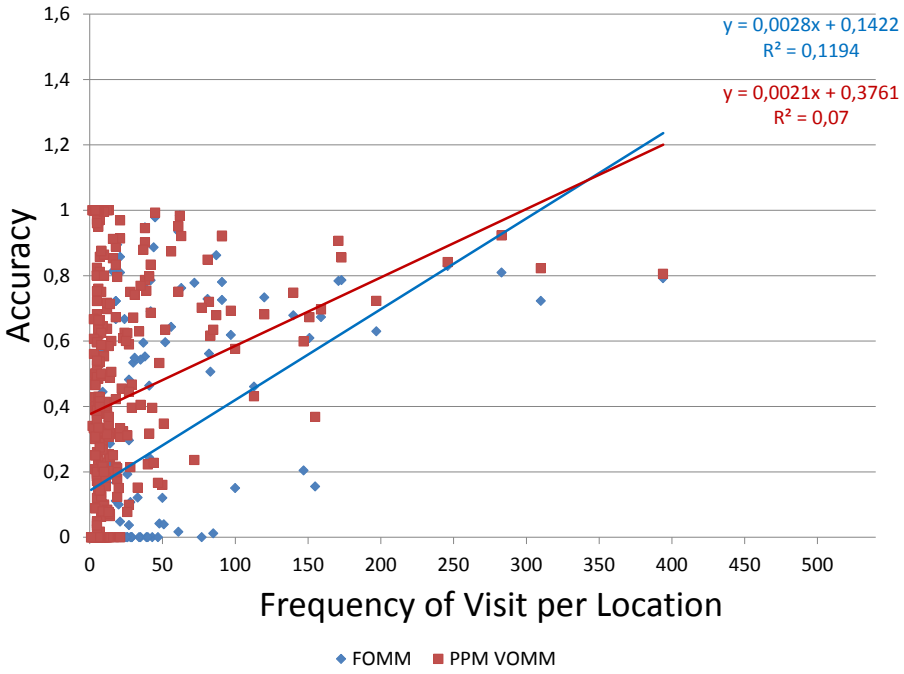

Figure 2.12: A positive correlation between frequency of visit per location and prediction accuracy exists for both FOMM ($r = 0.35, \rho = 0.32, P(\epsilon) = 0.0$) and PPM VOMM ($r = 0.26, \rho = 0.15, P(\epsilon) = 0.0034$).

FOMM as well as PPM VOMM depend to a high extent on the frequency of visit per location. This high dependency is confirmed by the strong correlation between frequency of visit per location and prediction accuracy, the value of both correlation coefficients are $r = 0.35, P(\epsilon) = 0.0$ and $\rho = 0.32$. Figure (2.12) demonstrates the positive correlation, PPM VOMM outperforms FOMM significantly for locations with a low frequency of visit. The accuracy of both models becomes similar as the frequency of visit per location increases.

Figure (2.13) shows the relationship between frequency of visit per location and improvement in accuracy. Improvement in accuracy decreases as the frequency of visits per location increases. This negative tendency is confirmed by both negative correlation coefficients ($r = -0.11, \rho = -0.003, P(\epsilon) = 0.79$). A two-sided unpaired t-test confirms the significance of the improvements $3 * 10^{-17}$. We conclude from the above results that PPM VOMM can detect patterns with a low frequency and thus can detect more patterns in the location history of a user.

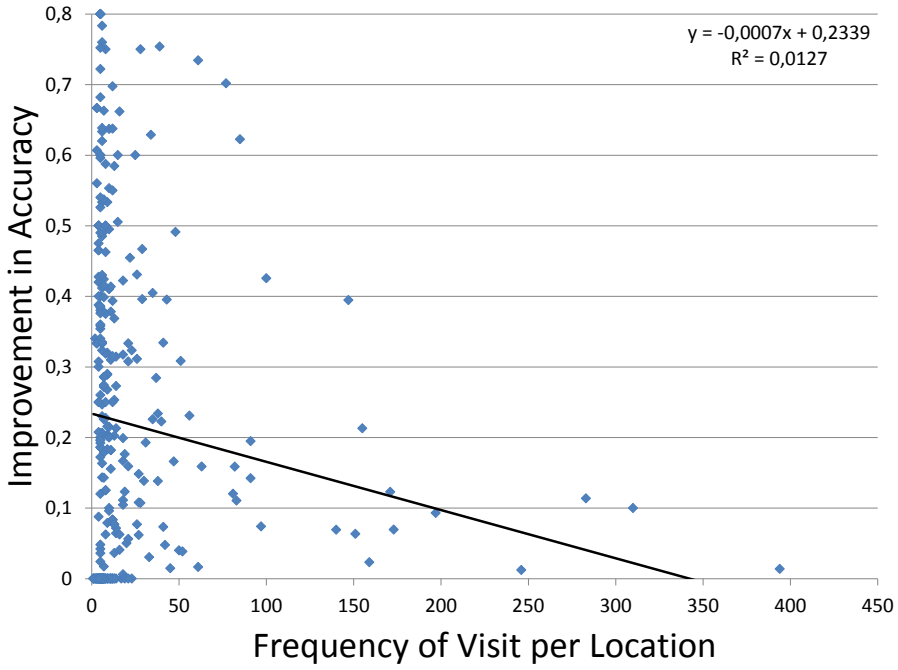

Figure 2.13: There is a negative correlation between frequency of visit per location and absolute improvement in accuracy ($r = -0.11, \rho = -0.003, P(\epsilon) = 0.79$).

2.7.3.6 Entropy

The two previous measurements illustrate the mobility predictability when taking both number of locations and the frequency of visits per location into account. Despite a high number of visited locations or a low frequency of visits per location, the mobility of a user might still be predictable to a considerable extent, for example when the location history of the user contains a few dominant locations with a very high frequency of visit and many locations with a very low frequency of visit. In order to combine both previous measurements we calculate the Shannon entropy for each user. The Shannon entropy quantifies the measure of uncertainty in the probability distribution of a random variable [Russell and Norvig, 2010, Page 703].

The Shannon entropy for a user whose location history contains many locations with similar frequencies is high, which means that the uncertainty of predicting the future location of the user is high, since no location has a significantly higher probability than the others. Indeed, the prediction accuracy of both FOMM and PPM VOMM negatively correlates with the entropy values to a very high extent. Both correlation coefficients $r = -0.74$ and $\rho = -0.69$ confirm the very strong negative correlation. Figure (2.14) plots both prediction accuracy and entropy. It clearly shows that PPM VOMM outperforms FOMM for all entropy values, with the tendency increasing as the entropy increases.

Figure (2.15) plots the improvement in accuracy with entropy. The plot demonstrates the tendency of improvement in accuracy to increase as entropy increases.

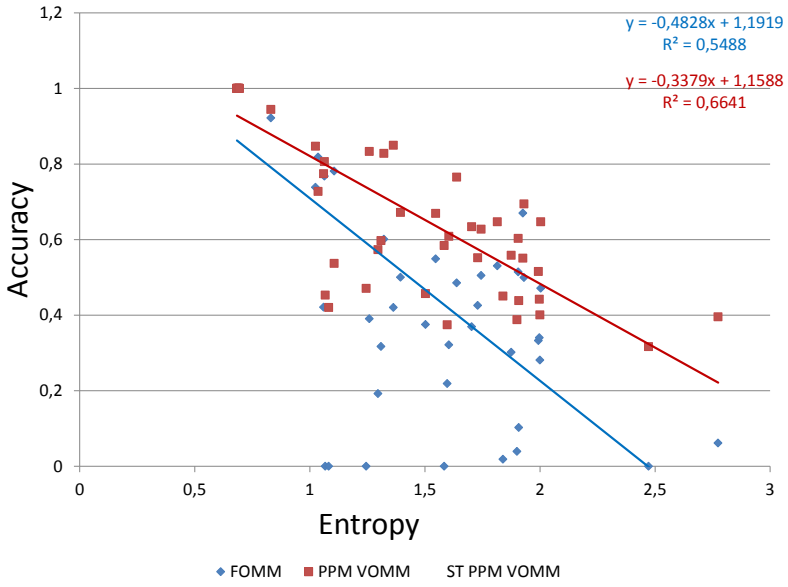

Figure 2.14: A negative correlation between entropy and accuracy exists for both FOMM ($r = -0.74, \rho = -0.69, P(\epsilon) = 0.0$) and PPM VOMM ($r = -0.86, \rho = -0.80, P(\epsilon) = 0.0$).

Both correlation coefficients $r = 0.43$ and $\rho = 0.50, P(\epsilon) = 0.0004$ clearly confirm this tendency. The above results repeatedly confirm the advantages of PPM VOMM compared to FOMM. PPM VOMM can better predict the mobility of highly entropic users with an explorative nature.

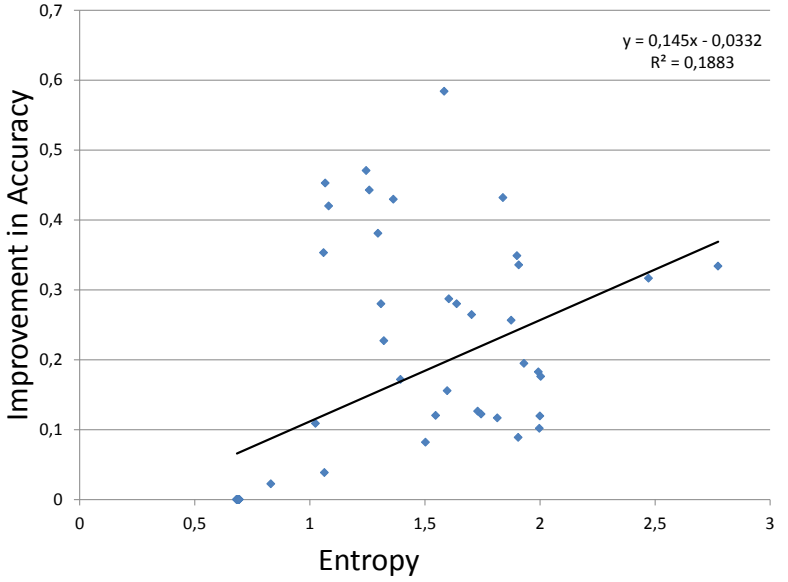

Figure 2.15: The plot shows a positive correlation between entropy and absolute improvement in accuracy ($r = 0.43, \rho = 0.50, P(\epsilon) = 0.0004$).

Chapter 3

Improving Location Prediction based on Temporal Context

As we have seen in the previous chapter, the location history of a user contains many spatial patterns, but many mobility patterns depend on both location and time. The spatial context helps detect frequent spatial pattern in the location history of the user, the inclusion of temporal context helps detect the returning intervals of a frequent spatial pattern. For example, a user might visit every Tuesday and Thursday evening after leaving their working location a fitness studio, and in the remaining three working days they go home. Thus the spatial pattern "home - work - fitness" studio has two returning intervals, namely Tuesday and Thursday evening. We refer to the returning intervals with the periodicities of the spatial pattern. In terms of predicting mobility, periodicity refers to patterns that recur after a certain period of time elapses. Taking periodicities into account helps detect more sophisticated patterns in the location history of a user. The inclusion of temporal context in the mobility model helps increase the amount of pattern in the location history of the user, thus contributing to improving the prediction accuracy of the mobility model. Further, mobility patterns are subject to decay over time. Incorporating the decay of mobility patterns at the design stage reduces the need for retraining the mobility model.

Chapter summary: An introduction to spatial temporal and pure temporal pattern is provided in section I. Related works on periodicity topics and temporal pattern are presented in section II. Mining periodic pattern and methods for detecting periodicities are discussed in section III. Section IV focuses on the inclusion of temporal context into the extended PPM VOMM mobility model in order to detect spatial, temporal or spatial temporal pattern. The performance of the extended PPM VOMM model is evaluated in the last section V of this chapter.

3.1 Introduction

The mobility model provided in the previous chapter (2) is able to predict the future location of an individual by focusing solely on the detection of frequent spatial patterns of certain length. The length of each pattern is less than or equal to the order of the underlying PPM VOMM mobility model. Focusing exclusively on uncovering frequent spatial patterns hampers the prediction of patterns with low frequency. Human mobility clearly exhibits temporal regularity, for example, we go home every evening, regardless of our current or previous locations. Consideration of temporal context and periodicities is of eminent importance in understanding and predicting human mobility. Many patterns occur at a certain time interval, beyond which they are not valid and therefore have a lower probability. Consideration of temporal context allows the detection and prediction of more patterns thereby increasing the accuracy of the mobility model.

The patterns may contain features related to the spatial or temporal contexts or both. Both time and space allow the detection of various types of pattern. The following pattern types can be distinguished:

- **Pure spatial pattern** is a pattern depending only on the sequence of locations visited by the user. Many locations are visited in a certain order, independently of the time, for example in sightseeing scenarios or at a museum, the order of visiting the various exhibits is always the same. A further example would be a walk in the park is always followed by a visit to the cafe before going home.
- **Pure temporal periodic pattern** is a pattern that depends only on time, for example a user goes home every evening regardless of their current location.
- **Spatial-temporal periodic pattern** is a pattern depending on both time and space. For example, a student visits the cafeteria at lunch time when they are at the university, i.e. the visit to the cafeteria is dependent on both spatial (university) and temporal (lunch break at 12:00 o'clock) context.

People are explorative in nature. They love to discover and visit new locations whenever they have the opportunity to do so. Therefore, they often visit new locations, places where they have never been before. Predicting a visit to a new location is very difficult, but it also complicates the prediction of further locations after this visit, because a new location corresponds to n unseen patterns according to the order of the underlying mobility model. Detection of pure temporal pattern helps to predict the future location of a user, even if their current location has not yet been seen in their location history. Furthermore, the detection of a spatial-temporal periodic pattern helps to predict non-frequent patterns such as attending a yoga course every Friday evening after work.

Frequent patterns are usually easy to detect and mostly less interesting and trivial such as "home - work - home" pattern. The detection of less frequent patterns is more interesting for service providers (Recommender Systems, intelligent advertising agencies) such as visiting certain shopping malls, locations associated with a

certain activity etc. Thus the utility of detecting non-frequent patterns is considerably higher [Ye et al., 2009], hence taking temporal context into account is of great importance for the utility of next location prediction.

Periodic Pattern Mining goes beyond considering only the order of visits or the time which elapses between two consecutive visits. Periodic Pattern Mining provides a method for detecting the periodicities of repetitive events such as a frequent spatial pattern. This chapter focuses on improving the prediction accuracy of our PPM VOMM model by detecting periodic patterns of the above types and integrating them into the PPM VOMM mobility model.

3.2 Related Work

The detection of periodic pattern has been the subject of many studies. Below we present a few related works.

3.2.1 Arrival, Stay & Departure Times

An investigation into the check-in behavior of users based on large scale LBSN data from an online LBSN platform called Foursquare with more than 12 million check-ins, 700 thousand users and a time period spanning over 100 days conducted in [Noulas et al., 2011a]. Users "check in" at locations using an application on their smartphones for example, by selecting a location from a list of locations the application locates nearby their current location. The investigation demonstrates how to reveal a spatial-temporal pattern. The analysis shows that the temporal distance between two consecutive check-ins varies; 10% of Foursquare check-ins occur within 10 minutes, 30% within 100 minutes and 20% within more than 2000 minutes. Further the authors found that 20% of the check-ins occur within a spatial distance of 1 km, almost 60% within 1-10 km and 20% within a spatial distance greater than 10 m. Both spatial and temporal distances between consecutive check-ins are important for uncovering any conditional dependencies between the locations of the two check-ins. For example a high temporal is an indicator that the two check-ins are not really dependent on each other.

The spatial and temporal mobility behavior of mobile users is investigated in [Chon et al., 2012]. The authors found that users usually spend $85 \pm 3\%$ of their time staying at particular locations and in the remaining time they transit between those locations. Users usually spend most of their time at a few frequent locations, for example they spend on average $83 \pm 12\%$ of their stay time at their top two most frequented locations and almost $94 \pm 3\%$ at their frequently visited locations. On average the users visit between 8 and 35 locations with a minimum frequency of three visits, these locations build only $11 \pm 2\%$ of the total number of locations visited by them. Further, the transitions of the users are mainly to their frequently visited locations, more precisely $69 \pm 7\%$ of their transitions are to frequently visited locations. The authors found that the duration of a stay at a location correlates positively with the arrival time at the current location, whereas the tendency to return to previously visited locations depends on the degree of spatial memory of

the model, i.e. the sequence of n last visited locations. Thus the mobility of users is highly predictable when both the spatial and temporal context of their mobility are taken into consideration [Chon et al., 2012].

3.2.2 Life Pattern

Life Pattern is a mining framework proposed in [Ye et al., 2009]. The framework is used for mining the individual life style and regular behavior of a user based on raw GPS data[Ye et al., 2009]. The mining framework uses the notion of a life pattern normal form (LP normal form) for a formal description of regularities in the life style of an individual. Significant locations of an individual play a central role in mining the life patterns of that individual. The mining framework therefore comprises of two phases, namely a pre-processing phase for transforming the raw GPS data into sequences of significant locations and a mining phase which uses the output of the first phase for detecting regular life patterns in the movement trajectories of an individual. The author distinguishes between different types of life pattern. An atomic life pattern consists of one single significant location. A sequential life pattern consists of a sequence of atomic life patterns such as "home - work - shopping mall - home". A temporal life pattern considers both arrival and stay time in an atomic pattern. An example of a temporal life pattern could be the frequency of cases where a user arrives at 10 o'clock somewhere and stays for two hours. A conditional life pattern contains both spatial and temporal constraints. A life associate rule combines two conditional life patterns which may be defined by rules such as, "A user arrives at a certain location within a certain time period with a certain probability whenever another conditional life pattern is true (such as taking a certain route)". The mining framework is able to detect a lot of patterns in the location history of an individual. The detected patterns allow the prediction of the future behavior/activity of an individual.

3.2.3 Conditional Random Fields

Significant location detection of a mobile user based on GPS traces and activity inference is the subject of a work by [Liao et al., 2007a]. The authors make use of temporal features such as divisions of the day (morning, afternoon, evening, etc.), the day of week and the duration of the stay, as well as a geographical database that contains the street map and locations near each street segment, and finally both velocity and GPS traces of each user in order to infer the activities of that user. The authors distinguish between two types of activity, namely low level activity such as driving, walking, taking a bus, etc. and significant activity such as leisure, shopping, visiting, working, sleeping, etc. They cluster the significant activities of the user to location types and use these location types for detecting the significant locations of the user. Both significant activities and locations of the user are inferred using a statistically orientated mining algorithm, namely conditional random field (CRF) [Liao et al., 2007a]. The proposed method of significant location detection is independent of a priori knowledge about minimum stay time or the number of significant locations. The authors claim that their method is more accurate than other methods such as K-means or DBSCAN, although they evaluated their method

based on GPS traces of four users collected over a period of one week. [Liao et al., 2007a].

3.2.4 Eigenbehaviors

[Eagle and Pentland, 2009] has proposed a statistically orientated mining framework based on Principal Component Analysis (PCA) for inferring the daily routines of users. The authors call the inferred routines the Eigenbehaviors of the user [Eagle and Pentland, 2009]. Formally, Eigenbehaviors are eigenvectors of the covariance matrix of behavior data [Eagle and Pentland, 2009]. The framework is evaluated using the Reality Mining dataset [Eagle and Pentland, 2006]. Among other location information (sequences of GSM cell tower IDs), the dataset contains the call logs of mobile phones, Bluetooth observations, etc. long-term social context is gathered in the form of a socio-matrix of relationships among the users and additional information from questionnaires.

The behavioral dataset is transformed into a binary matrix of 113 vectors. Each vector of the matrix represents a day containing 120 binary values for the 24 hours of the day multiplied by five states, namely $\{home, elsewhere, work, nosignal, off\}$. The dimension of the vector (120) contributes to the dimension of the patterns in the dataset. The n eigenvectors that have the highest eigenvalues are determined from the binary matrix based on PCA. The eigenvectors represent the user's set of primary eigenbehaviors. The individual behavior of a user can be approximated by a weighted sum of their primary eigenbehaviors.

The authors showed that 90% of the daily behavior of students can be characterized by few eigenvectors ("eigenbehaviors"). For each user they calculated the most populate eigenbehaviors in order to build the behavioral space of the user. Given the behavioral space and the behavior of the user in the first 12 hours of a previously unseen day, the behavior of the user within the remaining 12 hours can be predicted with an average overall accuracy of roughly 79%. Furthermore, generalities and group affiliations can be detected by combining the daily behavior of multiple users. As stated earlier, the study distinguishes five very coarse states (work, home, else, no signal and off) and three groups of students. It is questionable as to whether the model would accomplish the same level of accuracy using finer grained states and applied to users other than students.

3.2.5 Next Place

NextPlace is a prediction framework based on a non-linear time series analysis proposed by [Scellato et al., 2011a] and is a location independent predictor. For each location l_i in the model it analyzes the arrival times of the day A_i (i.e. the hours of day) and stay times R_i of each visit to that location l_i. Nextplace calculates a tuple of the form l_i, a_i, r_i for each location, where a_i is the average arrival time and r_i is the expected average stay time for location l_i. The leaving time for a location can be simply calculated by adding $a_i + r_i$. Once l_i, a_i, r_i are determined for all significant locations, NextPlace can predict the future location of the user at each time step t_i given Δt seconds as follows:

1. If the condition $a_i < t_i + \Delta t < a_i + r_i$ holds then location l_i is predicted. If more than one location holds the condition then a location is chosen randomly.

2. If the above condition is not satisfied for any location then the framework checks whether the condition $a_i < t_i + \Delta t$ holds. If yes, the algorithm is started again using $k+1$ most recent visits until k exceeds a threshold. Otherwise the predictor returns nothing, which means that the predictor believes that the user is not at a significant location.

The performance of NextPlace was tested based on four datasets. A location is considered to be significant for a user u if the user visits the location with a frequency of at least 20 visits. The model can predict the location of a user with an accuracy of 90%. The high accuracy of NextPlace is explained by the fact that only very frequently visited locations are considered (20 times) making only short to mid-term predictions ranging between a few seconds to one hour. Very frequently visited locations can be predicted without any great effort, especially for short to mid-term predictions. An extension of NextPlace has been proposed by [Vu et al., 2011] where more temporal features are considered such as day types (i.e. weekend or work day) instead of only considering the hours of the day.

3.3 Periodic Pattern Mining

The influence of time and the detection of periodic pattern is an important task of (temporal) data mining and play an important role in solving problems and improving the quality of different services from various application domains. [Mooney and Roddick, 2013] provides a comprehensive overview of different algorithms for mining periodic patterns in spatial-temporal data. The following subsections present a few related works on this topic.

3.3.1 Mining Sequential Patterns

Given a large dataset D of sequences of the form $s_1 s_2 ... s_n$, where $s_j = i_1, i_2 ..., i_m$ is an item set with m items, [Agrawal and Srikant, 1995] and [Srikant and Agrawal, 1996] have proposed methods for finding frequent patterns. A sequence s_j is said to be an $m - sequence$, i.e. it is of length m. A sequence $s_j = a_1, a_2, ..., a_n$ is contained in another sequence $s_k = b_1, b_2, ..., b_m$ with $n \leq m$ when $\forall a_i \in s_j \exists b_i \in s_k$ where $a_j \subseteq b_i$ and $j \leq i$. A sequence is maximal if it is not contained in another sequence.

The generalized sequential pattern (GSP) is an algorithm proposed in [Srikant and Agrawal, 1996] using a sub-string tree structure to store all possible sub-patterns of a maximal sequence and provide a counter to each node of the tree. Each node counter indicates the frequency of the sub-pattern associated with that node. The algorithm makes multiple passes over the dataset as follows. At the beginning, a set of frequent $1 - sequence$ is generated, where the frequency of each $1 - sequence$ exceeds a given $min_support$. Each subsequent pass uses the frequent sequences found in the previous step as a seed set and generates new candidate frequent sequences by adding one more item to them. The candidate sequences with a frequency greater

than $min_support$ build the seed for the next pass. The algorithm terminates when no frequent sequences can be found or when no further candidates can be generated. The method works very efficiently and scales well in time $O(mlogm)$ with m being the number of items.

3.3.2 Temporally Annotated Sequences (TAS)

A sequential pattern mining algorithm has been proposed by [Giannotti et al., 2006], which additionally considers the time elapsing between two consecutive item sets in a sequence in order to temporally annotate the two item sets such as:

$$A \xrightarrow{3.1} B \xrightarrow{9.4} C$$

Where A, B, C are item sets with one or more items, the numbers above the arrows represent the time elapsed between two consecutive item sets. The Temporally Annotated Sequence ($\mathfrak{TAS}$) is used to detect frequent sequences in a dataset based on $\tau - Containment$. A sequence T_1 is $\tau - Contained$ in another sequence T_2 if each item set $s_i \in T_1$ from the sequence T_1 is completely contained in an item set $s_j \in T_2$ of sequence T_2 with $i < j$ and the difference between two corresponding temporal annotations is less than a threshold τ. T_1 is $\tau - contained$ in D ($T_1 \preccurlyeq D$) if $T_1 \preccurlyeq T_2$ for some $T_2 \in D$. Given a set D of $\mathfrak{TAS}$, a time threshold τ and a minimum support threshold $s_{min} \in [0, 1]$, T_1 is frequent in D if it is $\tau - contained$ in at least a portion of all $T^* \in D$ greater than the threshold s_{min} ([Giannotti et al., 2006]).

3.3.3 Partial Periodic Pattern

The focus of Partial Periodic Pattern (PPP) is on identifying periodic sub-patterns in the dataset instead of full cyclic periodic pattern [Han et al., 1999, Rasheed, 2011, Han et al., 1998, Sheng et al., 2006]. For example in a period of one week, a user may visit a yoga course every Friday evening, but for the rest of the week may exhibit no movement regularity. Algorithms that treat the dataset as an inseparable flow of events (such as Fast Fourier Transform) do not allow the detection of partial periodic patterns [Han et al., 1999].

Let Σ be the set of features that can be derived from the dataset at any instant (for example, a set of temporal features), further let $*$ be a "don't care" symbol that can stand for any feature in Σ. A pattern $s = s_1, s_2, ..., s_p$ is a non-empty sequence of features with $s_i \in \Sigma \cup *$. The length of the pattern $|s|$ denotes the period of the pattern. A pattern s is called a i-pattern if it contains the letters from Σ at i positions. A sub-pattern $s' = s'_1, s'_2, ..., s'_p$ of $s = s_1, s_2, ..., s_p$ is a pattern with $s'_i \subseteq s_i$ for each $s_i \neq *$. The confidence $conf_s$ is the number of occurrences of s in the dataset divided by the total number of periods of length p. A frequent sub-pattern s' of s is a partial periodic pattern if its confidence exceeds a given threshold $conf_{min}$.

The max_sub-pattern tree hit set (MTHS) presented in [Han et al., 1999, Han et al., 1998] is an algorithm for finding all frequent partial periodic patterns in a candidate maximal sub-pattern C_{max} of period p. The authors extract c_{max} from

the set F_1 of all frequent 1-patterns. The root of the MTHS is set to c_{max}, each node at level $i+1$ has one non-* letter missing from its parent node at level i. $MTHS$ contains all sub-patterns of C_{max}. The set of sub-patterns with a confidence value greater than $conf_{min}$ represents a frequent partial periodic pattern in the dataset.

DPMiner is a further partial periodic pattern mining algorithm that uses a density based approach for identifying the 1-patterns (for period p) required to prune the search space [Sheng et al., 2006]. Beside $conf_{min}$ the algorithm uses an additional parameter d_{max} for identifying fragments with high density reachable occurrences of a letter $l_i \in \Sigma$. Two occurrences of the letter l_i are considered density reachable if the distance between the positions of the letter in both occurrences is less than d_{max}. Once the dense fragments are identified then the minimum period of the fragment, whose occurrence exceeds $conf_{min}$, can be calculated. The minimum period helps to prune the dense fragments by considering only those fragments with a minimum period less or equal to the period p. The dense region of the letter l_i is simply the union of all its pruned dense fragments. A $k-pattern$ can be mined by intersecting the dense regions of all letters in its item sets [Sheng et al., 2006].

3.4 The Inclusion of Temporal Features

Human mobility behavior obeys temporal regularities. Humans visit most of their locations within defined time periods. The time span between consecutive visits of a user to a location defines the periodic pattern according to which the user returns to the location. A periodic pattern can be fully cyclic, for example, if we consider a period of one day, a user goes home every evening. A periodic pattern can also be partial (a sub-pattern of the full period), for example, considering a period of one week, the movement behavior during work days and weekends differs significantly, a user goes to their work place every morning on work days but not on weekend days. Equally, the movement behavior of a user may differ on specific days of the week, perhaps even during certain hours of a certain weekday, for example a user visits a Spanish course after leaving work at location l every Tuesday and Friday afternoon.

The temporal context of an observation consists of n temporal features. Including temporal features helps detect less frequent patterns which occur during time periods which are different from the time period of the frequent spatial pattern depending on the current spatial context. For example, in the previous example, a mobility model that does not consider temporal features will only predict frequent patterns of the user and will fail to predict the location of the Spanish course. The model will probably predict the most frequent location after work, i.e. the home location of the user.

The movement behavior of human beings follows natural (day, week, month and year) and cultural (weekend) periodic patterns. These periods are structured hierarchically and match the structure of the PPM VOMM tree. The universal validity of these periodicities (at least in Western societies) is generally accepted and does not need further proof. Therefore we exploit these periodicities and integrate the temporal context in our PPM VOMM mobility model. The hierarchical structure of humankind periods, and the variable order of PPM VOMM allow detection of

temporal regularities and periodicities on varying scales (e.g. monthly, weekly and daily routines). Further, it allows pure temporal pattern to be detected when the current spatial context of the user does not occur in their history, hence improving the accuracy of the mobility model by predicting locations that obey only periodic patterns such as going home every evening.

3.4.1 Spatial-Temporal PPM VOMM (ST PPM VOMM) Tree

The alphabet Σ of the PPM VOMM model in the previous chapter contains only locations. In order to include temporal context we modify the tree and expand each spatial node with a sub-tree formed by the temporal features determined from the timestamps of the occurrences of the context s corresponding to that node. Each node is labeled with a temporal annotation from the temporal feature corresponding to that node. Let $\{F1, F2, ..F_i\}$ be the set of temporal features to be used. The alphabet of the ST PPM VOMM model is the union of both spatial locations and temporal features $\Sigma = \Sigma_{loc} \bigcup \Sigma_{temp}$ where Σ_{loc} are the location symbols and $\Sigma_{temp} = \bigcup_{j=1}^{i} F_j$ the domain of all temporal features. Table (3.1) shows the features included in the ST PPM VOMM model.

Variables	Domain	Description
Σ_{loc}	$\{l_1, l_2, ..., l_i\}$	The set of locations visited by the user
W	$\{Wd, We\}$	A binary variable representing whether it is a weekend day or a working day
D	$\{Sun, Mon, ..., Sat\}$	The day of week
$S^{\Delta t}$	$\{S_1, S_2, ..., S_j\}$	The number of time slots calculated by dividing the hours of day by Δt, setting $\Delta t = 1$ means the each hour of day represents a slot

Table 3.1: The features included in the ST PPM VOMM model, W, D and $S^{\Delta t}$ build together Σ_{temp}.

The extended model uses an inclusion semantic for implementing conceptual specificity, which allows the temporal features to be ordered in a manner such that the temporal features build a hierarchy of the form $F_1 \sqsubseteq F_2 .... \sqsubseteq F_n$. According to this inclusion semantic, F_1 represents the least specific feature (coarsest possible period in the movement of the user) and F_n represents the most specific temporal feature. The temporal feature F_i subdivides the temporal feature F_{i-1} into finer grained partitions of time (e.g. a day is sub-divided into 24 hours), thus the temporal futures with a higher index i represent more specific periods than temporal features F_j with a lower index $j < i$.

We determine the temporal annotation $\lambda = (\tau_1, \tau_2, ..., \tau_i)_q$ with $\forall j : \tau_j \in F_j$ for each spatial symbol (location) q appearing in the training sequence. For example, using the temporal features in table (3.1), a temporal annotation could be $(We, Sun, S_3^{\Delta t}$, where $S_3^{\Delta t}$ is the third time slice of the day with Δt duration. Each node in the tree (corresponding to a spatial symbol q and a context path s) becomes a temporal sub-tree with nodes labeled with temporal annotations according to the temporal features used. Each node in the temporal sub-tree at level i has a label

from the temporal feature F_i and a counter indicating the occurrence of the sequence sq fitting the temporal feature indicated by the label.

Figure (3.1) depicts an exemplary temporal sub-tree for a spatial node. The root of the node represents a spatial node of the parent spatial tree. Each node of the temporal sub-tree is annotated with a temporal feature. Assuming the use of work or weekend day, day of week and hours of day as temporal features, a path represents a temporal annotation like "Work Day - Monday - S_{10}" for a visit to location q on Monday at ten o'clock. The temporal sub-tree for a location q that occurs only one time at Tuesday 4:00 pm. in the location history is a linear list like "(q, 1) - (wd, 1) - (tu, 1) - (S_{16}, 1)".

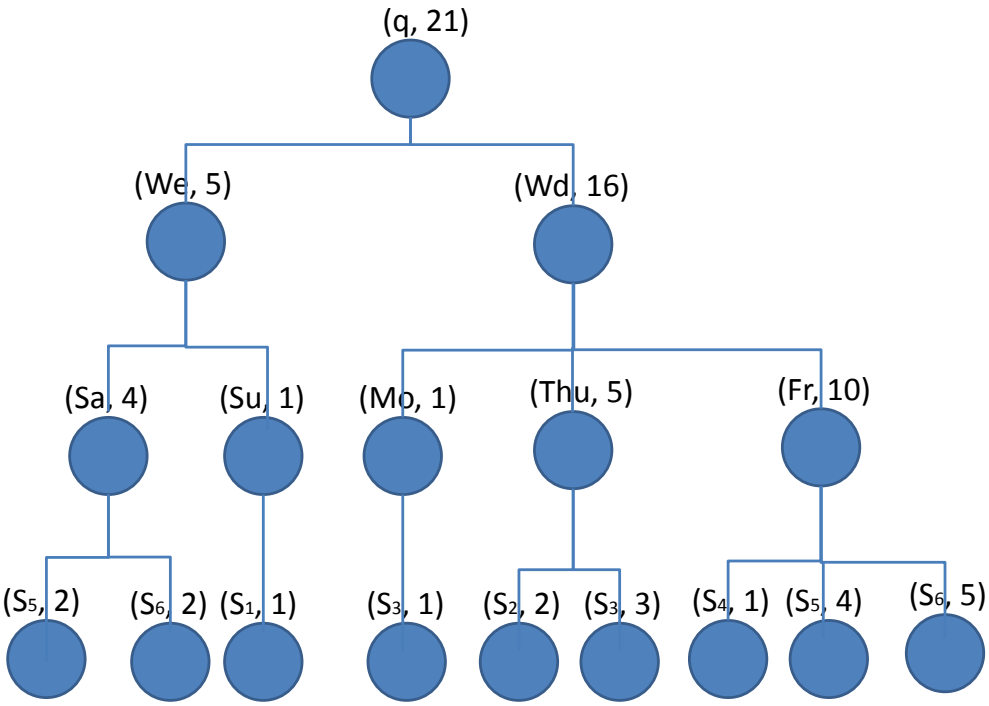

Figure 3.1: An example of a temporal context sub-tree. The root of the tree is spatial node, each node is annotated with a temporal feature, a path in the tree represents a temporal annotation like: Weekend - Saturday - 6 O'Clock. Each node has a counter c for bookkeeping the occurrence of location q at the time specified by the node

3.4.1.1 Training ST PPM VOMM

The update mechanism for the standard PPM VOMM remains the same, for example, given a spatial context sq and temporal annotations $(Wd, Mon, S_2^{(6)})$, all counters along the spatial-temporal path must be incremented in exactly the same way as in the PPM VOMM tree with one exception, if the stay time of the user at a location q spans over multiple time slices, all leaves corresponding to these time slices must be incremented.

3.4.1.2 Prediction Using ST PPM VOMM

The standard PPM algorithm can be used during the prediction phase, because the inclusion semantic of the temporal features allows the use of periods of various length without modification. Given a context s, a location q and temporal features $(Wd, Mon, S_2^{(6)})$ ST PPM VOMM algorithm performs as follows. Similar to PPM, ST PMM VOMM traverses the tree according to the spatial context s in order to find a child node labeled by q, if a path for the current spatial context s does not exist, the algorithm escapes to a lower spatial order until a node is found, or escapes to the node q under the root of the tree. Once a node is found, ST PPM VOMM traverses then the temporal subtree of that node according to the extracted temporal features. When no node can be found according to the specific periodicity, ST PPM VOMM escapes to a more general periodicity, for example, if no node can be found according to the temporal features $(Wd, Mon, S_2^{(6)})$, ST PPM VOMM escapes automatically to the more general periodicity given by the temporal features (Wd, Mon). The algorithm continues escaping until a period is found, or escapes to the pure spatial context given by the root node of the temporal subtree. The model predicts the next location based on pure temporal patterns if the current spatial context s is empty, because ST PPM VOMM then uses the temporal subtree of the node labeled by the location q (whose probability should be estimated) under the root of the spatial tree. Pure temporal patterns allow predictions even if the current location of the user is unknown, for example when the user visits a shopping mall for the first time after work, the pure temporal pattern helps detect the next location of the user, namely their home location. The full joint probability $p(q|s, \lambda)$ can be calculated according to equations (2.10) and (2.11) from chapter 2.

Next location prediction does not depend on fixed time intervals when predicting the future location of the user, because the duration of stay at a location depends on the arrival time at that location [Chon et al., 2012], which varies between a few minutes to few hours or sometimes even more than a day. Further, it has been shown in [Chon et al., 2012] that the tendency to return to a location depends solely on the spatial order of the mobility model. It is therefore reasonable to assume conditional independency between both temporal and spatial context s, the spatial-temporal probability $p(q|s, \lambda)$ can then be rewritten according to equation (3.1)

$$P(q|s, \lambda) = P(q|s) * P(q|\lambda) \tag{3.1}$$

where λ is the set of temporal annotations corresponding to the current occurrence of q. The probability $p(q|\lambda)$ can be calculated using the counters in the temporal sub-tree of node q under the root ϵ of the ST PPM VOMM tree (equation 3.2).

$$\tilde{P}(q|\lambda) = \frac{C(q\lambda'\tau_j}{|\Sigma_{q\lambda'}| + \Sigma_{\tau' \in \Sigma_{q\lambda'}} C(q\lambda'\tau')} \tag{3.2}$$

where $\lambda = \lambda'\tau_j$, for example, if λ contains "work day - day of week - hours of day", then λ' contains "work day - day of week" and τ_j represents the hours of day. ST PPM VOMM allows calculation of the full joint probability according to the equation (2.11) and calculation assuming conditional independency between both

spatial context s and temporal context according to equation (3.2). We evaluate which of the two equations achieves a better accuracy later.

3.4.2 Drift Function

The life of every individual contains significant events that have a great influence on their behavior. Some of these events cause changes in their long-term behavior, including mobility patterns. Examples of such events are marriage, a move, the birth of a child, a new job, a new circle of friends, etc. For these reasons, our mobility patterns are subject to change over time. Mobility models based on probabilistic reasoning are particularly vulnerable to these changes, because statistically it takes a long time until the probability of a new pattern is higher than the probability of a frequent pattern that has become invalid. During this time, the performance of the mobility model worsens significantly.

The adaptability of the mobility model to changing mobility behavior of users can be increased by using a drift function. With each non-occurrence of a mobility pattern, the drift function causes the mobility pattern to decay faster. The drift function has two parameters, the first parameter is a hyper-parameter α that controls the degree of drift and the second parameter controls the time unit Δt to which the mobility pattern should decay according to α. With each non-occurrence of a mobility pattern, the drift function causes this to decay to factor $1 - \alpha$ after each Δt time unit.

Let $t_i < t$ be the time of the last occurrence of the context s. Equation (3.3) shows the drift function, which calculates at any point in time the degree of drift of context s.

$$\varpi(s, t) = (1 - \alpha)^{(t-t_i)/\Delta t} \tag{3.3}$$

where s is a spatial context, α is a hyper parameter representing the degree of drift and Δt is the time unit after which the mobility pattern will decay. The unit of the factor $t - t_i$ is set to the most specific temporal feature used in the mobility model, namely Δt. $t - t_i$ is thus a multiple of Δt, the longer ago is the last occurrence of a context s, the higher its decay. For example, if we use a temporal feature that subdivides the time of the day into smaller time slices of 6 hours, then the counter is drifted every 6 hours by factor $(1 - \alpha)$. The value of $\varpi(s, t)$ is always within the interval $[0, 1]$.

Upon receiving a new observation, each corresponding node in the VOMM tree is incremented by multiplying the previous counter $C(s)$ of the node by the drift function and then incrementing it by one as shown in equation (3.4).

$$C'(s) = C(s) * \varpi(s, t) + 1 \tag{3.4}$$

Equations (2.11) and (2.12) must be modified in order to use the drift function. Equations (3.5) and (3.6) illustrate the modifications.

$$\tilde{P}(q|s) = \frac{C'(sq) * \varpi(sq,t)}{|\Sigma_s| + \Sigma_{q' \in \Sigma_s} C(sq') * \varpi(sq',t)} \tag{3.5}$$

$$\tilde{P}(escape|s) = 1 - \sum_{q \in \Sigma_s} \tilde{P}(q|s) = \frac{|\Sigma_s|}{|\Sigma_s| + \Sigma_{q' \in \Sigma_s} C(sq') * \varpi(sq',t)} \tag{3.6}$$

where $\varpi(sq,t)$ is calculated according to the last occurrence of context sq.

3.5 Empirical Results

The empirical results showed that the temporal feature D for the day of week has only a marginal effect on the accuracy of ST PPM VOMM, consequently we decided to use only the remaining two temporal features ($\lambda = (W,S)$) where $W = \{Wd, We\}$ indicates the type of day (i.e. whether it is a work day Wd or a weekend day We) and the temporal feature $S^{\Delta t}$ that partitions the day into time slices of the size Δt. We set $\Delta t = 6$, which partitions the day into four time slices(morning($S_1^{(6)}$), afternoon($S_2^{(6)}$), evening($S_3^{(6)}$) and night($S_4^{(6)}$)). All the following empirical results are based on the GeoLife dataset.

3.5.1 Conditional Dependency

The empirical results showed that the temporal context conditionally depends on the current location and not on the whole spatial context. The best accuracy of the ST PPM VOMM for full conditional dependency between the temporal and the whole spatial context was (66.2) with the parameter settings (($\delta = 10, \epsilon = 10, minPt = 5$)), order 2 and using 90% of the data as training data. However the model that assumes that temporal context only depends on the current location (according to equation 3.2) achieves a significantly higher accuracy of 68.9%, which corresponds to an absolute improvement of 0.027 and a relative improvement of 4.1%.

The accuracy of ST PMM VOMM is higher compared to both FOMM and PPM VOMM. The absolute improvement in accuracy of ST PMM VOMM when compared to PPM VOMM is around 0.073 and it increases to 0.269 when ST PMM VOMM is compared to FOMM, which corresponds to a relative improvement in accuracy of 12% and 58% respectively. The significance of the improvements in accuracy is confirmed by both two-sided paired and two-sided unpaired t-tests, shown in Table (3.2).

Dataset	T-Test type / Mobility Model	two-sided Paired	two-sided Unpaired
GeoLife	FOMM	$4.6 * 10^{-5}$	$9.3 * 10^{-5}$
	PPM VOMM	0.037	0.075

Table 3.2: The results of the t-test analysis for both two-sided paired and two-sided unpaired t-tests for showing the significance of the improvements in accuracies using ST PPM VOMM compared to FOMM as well as PPM VOMM.

3.5.2 Performance Analysis

In the following subsections we evaluate the performance of ST PPM VOMM by comparing it to the FOMM as well as the PPM VOMM models from the previous chapter.

3.5.2.1 Size of Training Data

ST PPM VOMM is as stable as PPM VOMM in terms of the size of training data. Table (3.3) shows the prediction accuracies of FOMM, PPM VOMM and ST PPM VOMM with varying amounts of training data. The accuracy of ST PPM VOMM drops off to only 1.2% when reducing the amount of training data from 90% to zero. Further, inclusion of temporal features causes ST PPM VOMM to detect more patterns in the same amount of training data, namely spatial, temporal and mixed spatial-temporal pattern. Thus, inclusion of temporal features further alleviates the effects of the cold start problem. ST PPM VOMM does not even require training data to achieve a higher prediction accuracy than FOMM and PPM VOMM both trained with 90% of the data.

Dataset	Mobility Model \ Training Data %	0.9	0.75	0.5	0.25	0.0
GeoLife	FOMM	52.01	48.29	43.72	39.13	0.0
	PPM VOMM	61.63	61.57	61.51	61.56	60.60
	ST PPM VOMM	68.9	68.7	68.9	68.5	67.7

Table 3.3: Prediction accuracies in percent achieved using different mobility models and varying amounts of training data. The order of all models is set to two.

3.5.2.2 Context Length (Order)

The amount of possible pattern in the location history of a mobile user increases when increasing the order of the underlying mobility model. Increasing the order of a model increases the demand for training data. The amount of training data correlates with the increasing order of the underlying mobility model. Increasing the order of the model without more training data causes the accuracy of the model to decrease.

Dataset	Mobility Model \ Order	2	3	4	5
GeoLife	FOMM	43.67	33.48	25.95	19.84
	PPM VOMM	61.51	61.85	60.54	59.98
	ST PPM VOMM	68.87	68.51	67.57	66.86

Table 3.4: Prediction accuracies in percent when varying the order of the underlying mobility model.

Table (3.4) shows the results of all three mobility models when the order of the mobility models is varied between two and five. Indeed, the accuracy of all models decreases as the order increases. The accuracy of FOMM drops off from 43.67% to 19.84% when the order of the model is increased from two to five. Similarly to the PPM VOMM, the accuracy of the ST PPM VOMM exhibits only a slight reduction of around 2%. The secret of the stability of both the PPM VOMM and ST PPM VOMM lies in their variable order, the models automatically switch to a lower order when no pattern can be found with the higher order. The escape mechanism of the ST PPM VOMM, the variable order and the ability of the model to detect spatial, spatial-temporal and pure temporal patterns causes the model consistently to achieve high prediction accuracy.

3.5.2.3 Drift Function

The drift function turned out to have a positive influence on the performance of the ST PPM VOMM. Setting the value of $\alpha = 0.007$ maximizes the accuracy of ST PPM VOMM to 70.15%. Using the drift function, the absolute accuracy is increased by 1.29%, which corresponds to 1.87% relative improvement in accuracy. Setting $\alpha = 0.007$ means that the influence of a pattern on the mobility behavior of a user disappears after a maximum of two years of non-occurrence.

3.5.2.4 Zero-Frequency

As discussed earlier in section 2.7.3.3, mobility models based on probabilistic reasoning suffer from the zero-frequency problem. Zero-frequency occurs when a new context is observed during the prediction phase. The inclusion of temporal features into the mobility model increases the amount of detected mobility patterns in the location history of the user, because in addition to spatial patterns, spatial-temporal as well as pure temporal patterns become detectable. The detection of more patterns at the same time means detection of more sub-pattern (due to the variable order of ST PPM VOMM). Hence, the prediction accuracy can be improved even if the amount of training data remains the same.

As already mentioned in the previous chapter the GeoLife dataset contains 23.73% new contexts, that were not seen during the training phase. ST PPM VOMM can correctly predict 61.18% (56.44% without the drift function) of the unseen contexts. This contribution corresponds to a total absolute improvement of 0.1467(58%) in accuracy compared to FOMM, and 0.031(42%) compared to PPM VOMM (the numbers in parenthesis represent percentages of total improvements in accuracy). This result shows that including temporal context in the mobility model alleviates the negative impact of the zero-frequency problem and thus contributes to significant accuracy improvements confirmed by two-sided unpaired t-test values of $9 * 10^{-24}$ when compared to FOMM, and 0.0045 when compared to PPM VOMM.

Figure (3.2) and (3.3) show the distribution of improvements due to alleviation of the impact of zero-frequency over the hours of work days compared to PPM VOMM and FOMM respectively. The size of each bubble indicates the amount of unseen context during that hour. Most of the unseen contexts occur during the morning

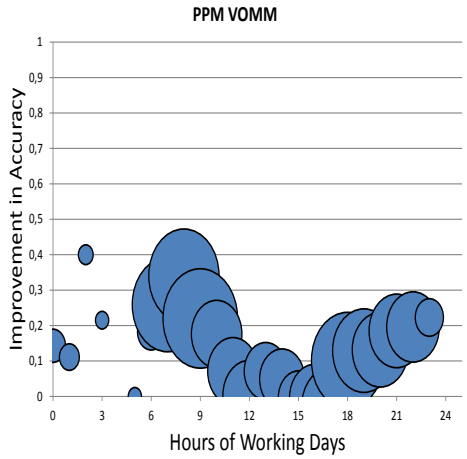

Figure 3.2: The distribution of absolute improvements in accuracy due to the alleviation of the effects of zero-frequency problem over the hours of work days compared to PPM VOMM.

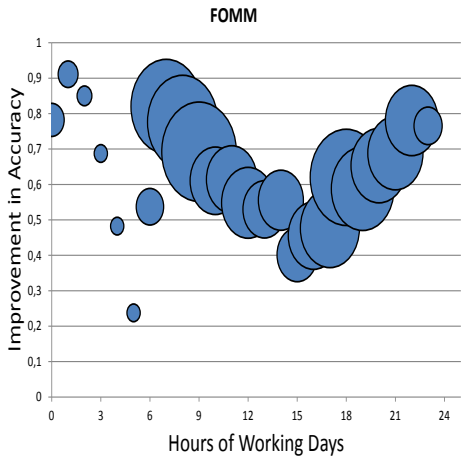

Figure 3.3: The distribution of absolute improvements in accuracy due to the alleviation of the effects of zero-frequency problem over the hours of work days compared to FOMM.

and evening hours before the users go to their most significant locations, namely their work places in the morning and their homes in the evening. In the morning, most improvement is achieved between seven and nine a.m. and drops off smoothly towards the noon hours. Intuitively, the probability of arriving at a work place decreases towards noon. Improvements in the evening increase smoothly beyond six p.m. towards the late night hours, because the later the evening, the higher the probability of going home. The distribution of improvements over the hours of work days suggests that the improvements are achieved due to detection of pure temporal pattern (independently from the current location of the user) such as going to the work place every morning between seven and nine a.m. or going home at some point in the evening, mostly between five and seven p.m.

Figure (3.4) shows the distribution of improvements in accuracy over the hours of weekend days compared to PPM VOMM. Unlike work days, unseen context generally occurs between 10-12 a.m., and during the evening hours between 8-10 p.m. The improvement in accuracy compared to PPM VOMM has two dominating ranges, the first range between 9-12 a.m., and the second range 6-10 p.m. The improvement in accuracy in the morning increases smoothly towards the noon hours and reaches its maximum around 12 a.m. The improvement in accuracy in the evening increases smoothly and reaches its maximum at around 9-11 p.m. The distribution fits the life of an adult at the weekend. During the morning hours, a person either goes shopping (mostly on Saturdays) or to a brunch (on Sundays) and then goes home afterwards. In the evening, the person visits restaurants or bars and during "homecoming" hours goes home. Once again we conclude that purely temporal patterns are responsible for the improvements.

Figure (3.5) shows the distribution of improvements in accuracy over the hours of weekend days compared to FOMM. The amount of unseen contexts and the improvements in accuracy increases smoothly towards the late night hours when the

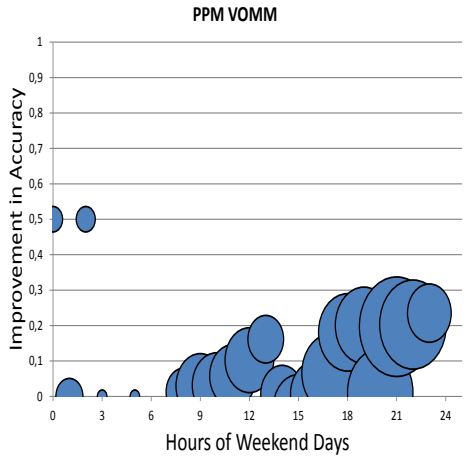

Figure 3.4: The distribution of absolute improvements in accuracy due to the alleviation of the effects of zero-frequency problem over the hours of weekend days compared to PPM VOMM.

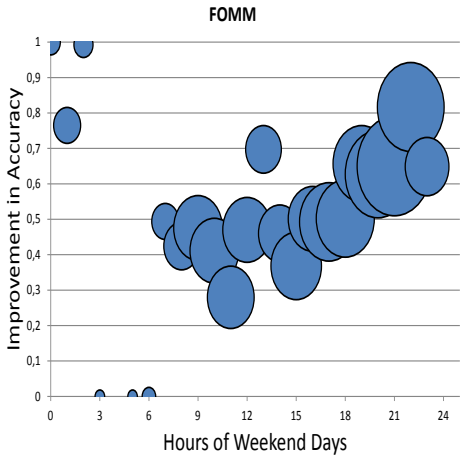

Figure 3.5: The distribution of absolute improvements in accuracy due to the alleviation of the effects of zero-frequency problem over the hours of weekend days compared to FOMM.

users are most active. It demonstrates again the importance of the variable order of ST PPM VOMM and the inclusion of temporal context for improving the accuracy during the hours of day when people are most explorative and hard to predict. Generally, ST PPM VOMM can alleviate the impact of zero-frequency during the night hours on weekend days and during the morning hours of work days to a high extent, i.e. by up to 85%.

3.5.2.5 Number of Locations

The relationship between the accuracy of ST PMM VOMM and the number of locations visited by the users exhibits the same negative tendency as FOMM and PPM VOMM. The accuracy of all models has a strong negative linear relationship, confirmed by the correlation coefficients on figure (3.6). The accuracy of ST PPM VOMM drops off when the number of locations increases, but to a lesser extent than that of FOMM or PPM VOMM. Figure (3.6) shows the relationship between the accuracy of all three models and the number of locations. It can easily be seen from the slope of the fitted lines that the accuracy of ST PPM VOMM drops off to a lesser extent. The linear dependency between the accuracy of ST PMM VOMM and the number of locations is 21%, lower than PPM VOMM and almost 39% for FOMM.

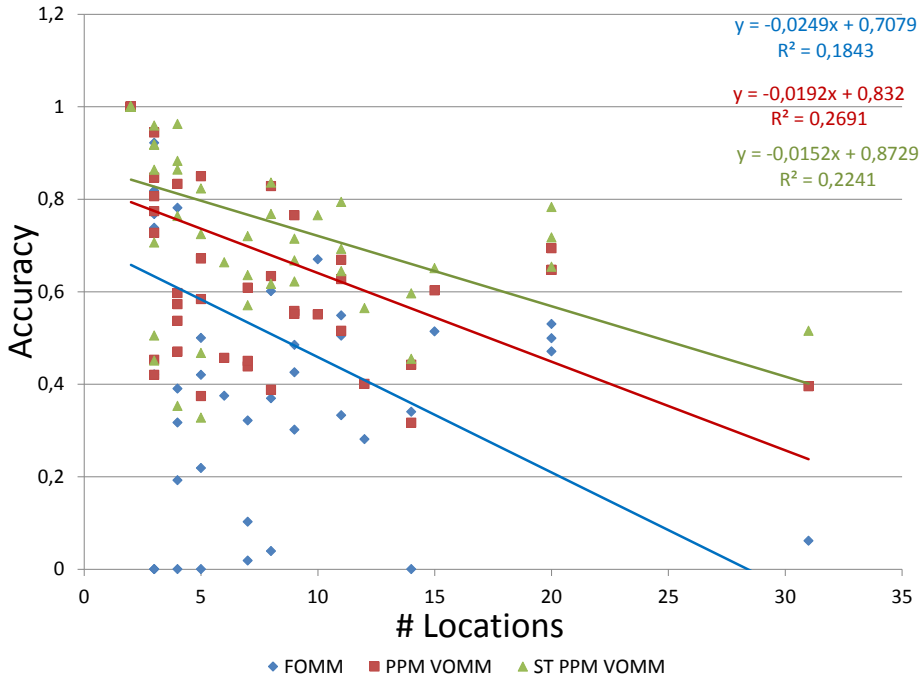

Figure 3.6: A negative correlation between the number of locations and the accuracy exists for FOMM ($r = -0.43, \rho = -0.59, P(\epsilon) = 0.0$), PPM VOMM ($r = -0.52, \rho = -0.69, P(\epsilon) = 0.0$) and ST PPM VOMM ($r = -0.47, \rho = -0.70, P(\epsilon) = 0.0002$).

Although the accuracy of all three models drops off when the number of locations increases, ST PPM VOMM has significant advantages over both FOMM and PPM VOMM. The improvement in accuracy due to the inclusion of temporal features exhibits a positive tendency with the number of locations (figure (3.7)). The improvement in accuracy compared to PPM VOMM exhibits a positive correlation confirmed by both $r = 0.22$ and $\rho = 0.50, P(\epsilon) = 0.0$. Both correlation coefficients are even higher for the improvements in accuracy compared to FOMM, which confirms a strong positive correlation with coefficient values $r = 0.30$, $\rho = 0.55 and P(\epsilon) = 0.0$. The above results show the importance of including temporal context in ST PPM VOMM, because it improves the prediction accuracy of the mobility of explorative users, despite a higher uncertainty due to visiting numerous locations.

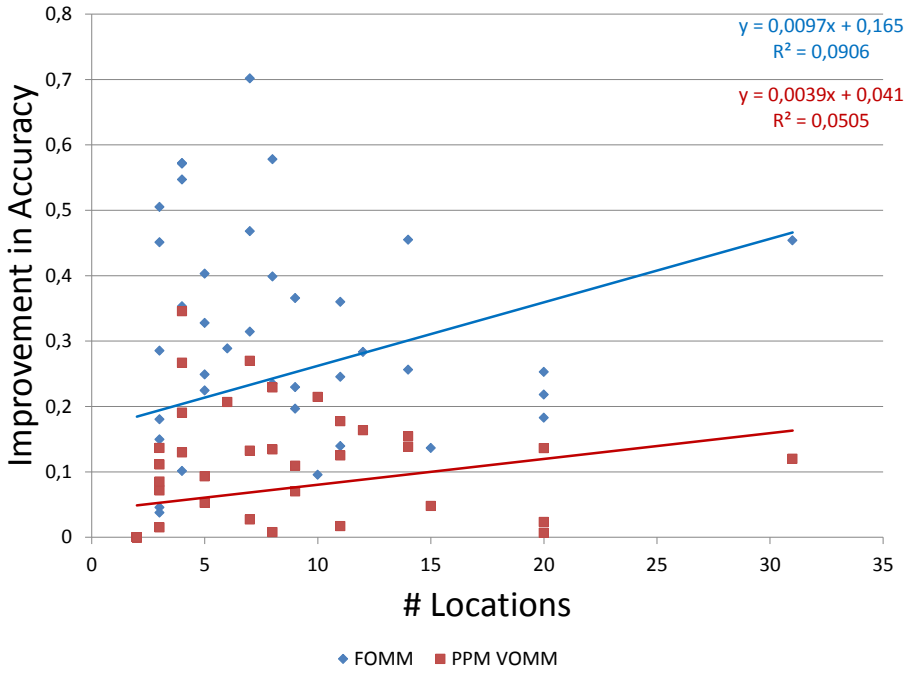

Figure 3.7: A positive correlation between the number of locations and absolute improvements in accuracy using ST PPM VOMM exists compared to both PPM VOMM ($r = 0.22, \rho = 0.50, P(\epsilon) = 0.0002$) and FOMM ($r = 0.30, \rho = 0.55, P(\epsilon) = 0.0$).

3.5.2.6 Frequency of Visit per Location

Similarly to PPM VOMM and FOMM; ST PPM VOMM depends to a very high extent on the frequency of visits per location. Accuracy increases as the frequency of visits increases. Figure (3.8) shows the relationship between accuracy and the frequency of visit per location for all three models. The accuracy of all three models shows moderate to strong positive correlations confirmed by both correlation coefficients. PPM VOMM achieves a higher accuracy for locations with a lower frequency of visits compared to FOMM, but it achieves accuracies similar to FOMM for locations with a higher frequency of visit. Unlike PPM VOMM, the inclusion of temporal context in ST PMM VOMM leads to a consistently higher improvement in accuracy for all locations, despite their frequency of visit. Thus ST PPM VOMM has clear advantages over both FOMM and PPM VOMM.

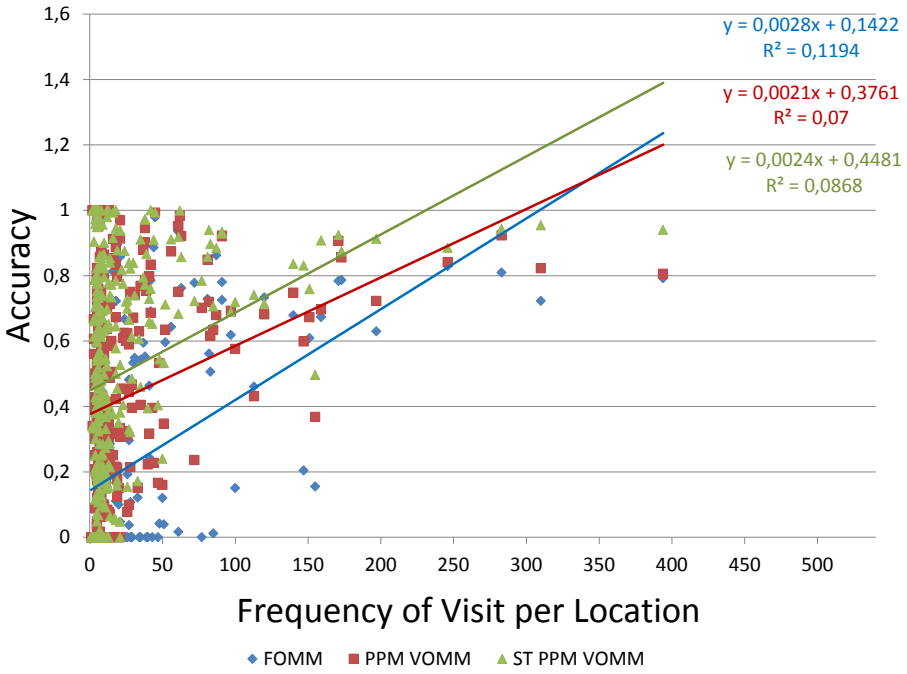

Figure 3.8: A positive correlation between the history size per location and the accuracy exists for FOMM ($r = 0.35, \rho = 0.32, P(\epsilon) = 0.0$), PPM VOMM ($r = 0.26, \rho = 0.15, P(\epsilon) = 0.0034$) and ST PPM VOMM ($r = 0.29, \rho = 0.21, P(\epsilon) = 0.0$).

Figure (3.9) shows the relationship between the improvements in accuracy and the frequency of visit per location using ST PMM VOMM compared with both PPM VOMM and FOMM. The improvements in accuracy compared to FOMM show a negative tendency, which is confirmed by slightly negative correlation coefficient values of $r = -0.06$ and $\rho = -0.10, P(\epsilon) = 0.0536$. The low correlation coefficient values indicate that ST PPM VOMM achieves a consistently high performance compared to FOMM, despite the frequency of visit. The improvements in accuracy compared to PPM VOMM show a positive tendency confirmed by a slightly positive correlation according to Pearson's correlation coefficient and a strong positive correlation according to Spearman's rank correlation coefficient. The inclusion of temporal context in ST PPM VOMM increases the low improvement in accuracy using PPM VOMM compared to FOMM, which again demonstrates the importance of temporal context in next location prediction.

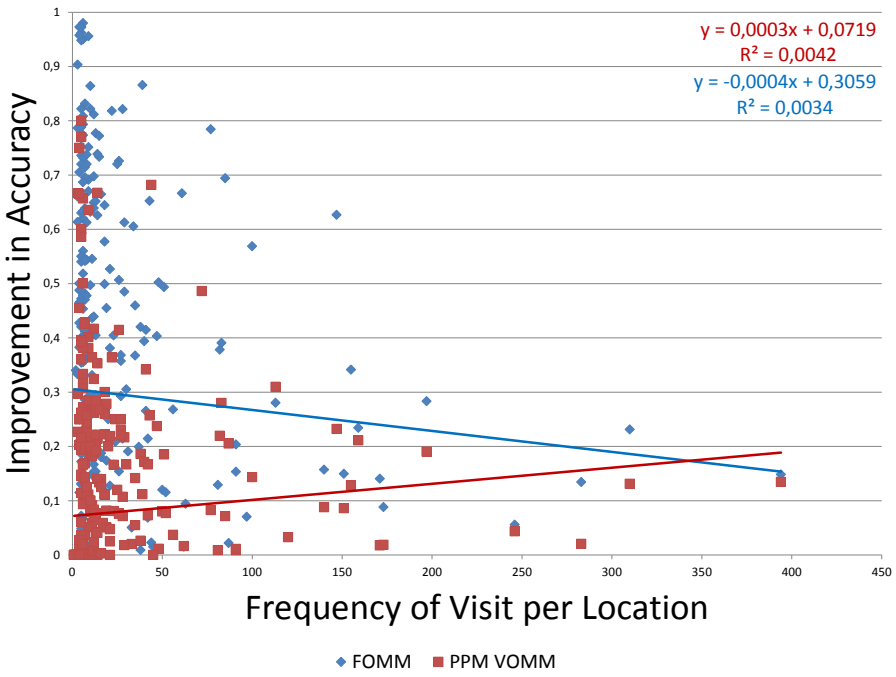

Figure 3.9: A positive correlation between the average history per location and absolute improvement in accuracy using ST PPM VOMM exists compared to both FOMM ($r = -0.06, \rho = -0.10, P(\epsilon) = 0.0536$) and spatial PPM VOMM ($r = 0.06, \rho = 0.29, P(\epsilon) = 0.0$).

In order to underline the significance of the improvements in accuracy achieved by the inclusion of temporal features in ST PPM VOMM compared to both FOMM and PPM VOMM we conducted a t-test analysis. The results shown in Table (3.5) underline the significance of the improvements compared to both FOMM and PPM VOMM using both two-sided paired and two-sided unpaired t-tests.

T-Test Type / Mobility Model	Two-Sided Paired	Two-Sided Unpaired
FOMM	$1.8 * 10^{-28}$	$3.6 * 10^{-28}$
PPM VOMM	0.0012	0.0024

Table 3.5: Frequency of visit per location: the results of the t-test analysis for both paired and unpaired t-tests regarding the significance of the improvements in accuracy using ST PPM VOMM compared to FOMM and PPM VOMM.

3.5.2.7 Entropy

The prediction accuracy of all three models statistically has a strong negative dependency on the entropy of user. That is, the mobility of highly entropic users is less predictable, because of the strong uncertainty associated with the next visit of the user. The strong negative dependency is underlined by a very strong negative correlation using both correlation coefficients $r = -0.73$ and $\rho = -0.76, P(\epsilon) = 0.0$. The

slopes of the lines for the three models in figure (3.10) show that the accuracy of ST PPM VOMM decreases more slowly by a factor of 22.5% compared to PPM VOMM and by a factor of 46% in comparison with FOMM. Thus the inclusion of temporal context contributes to a slower loss in performance as the entropy increases. The above results impressively demonstrate that ST PPM VOMM increasingly outperforms both FOMM and PPM VOMM with increasing entropy.

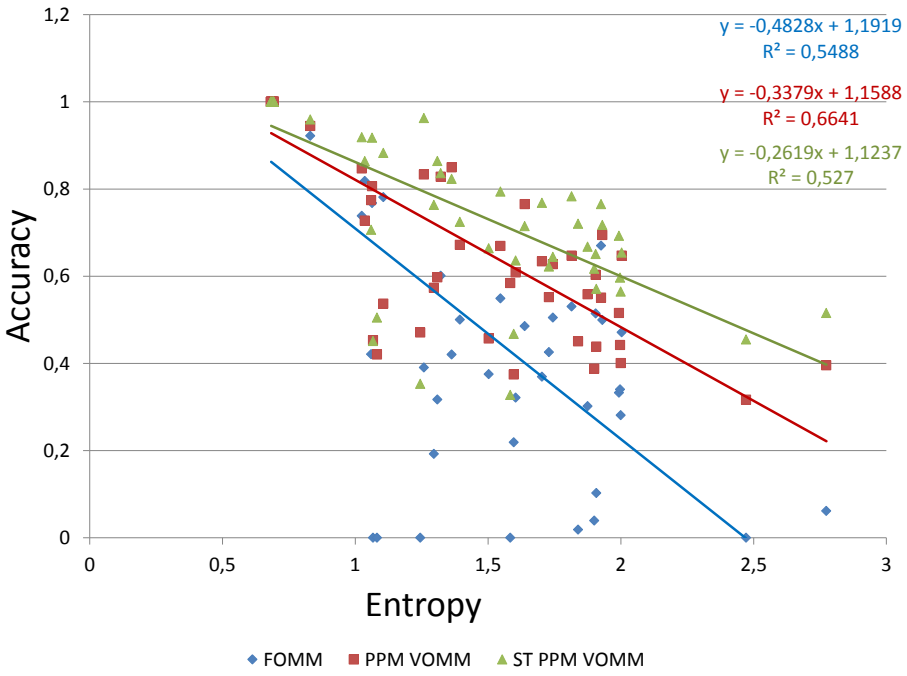

Figure 3.10: A negative correlation between the entropy and the accuracy exists for FOMM ($r = -0.74, \rho = -0.69, P(\epsilon) = 0.0$), PPM VOMM ($r = -0.86, \rho = -0.80, P(\epsilon) = 0.0$) and ST PPM VOMM ($r = -0.73, \rho = -0.76, P(\epsilon) = 0.0$).

Figure (3.11) plots the relationship between entropy and the improvements in accuracy using ST PPM VOMM compared to both FOMM (Blue line) and PPM VOMM (Red Line). Both lines show a positive tendency, the improvement in accuracy increases as the entropy increases. The tendency is statistically confirmed by a strong positive correlation compared to PPM VOMM ($r = 0.39, \rho = 0.55, P(\epsilon) = 0.0$) and a very strong positive correlation compared to FOMM ($r = 0.61, \rho = 0.63, P(\epsilon) = 0.0$).

Both the variable order and the escape mechanism of the standard PPM VOMM combined with the inclusion of temporal features increase the predictability of the mobility of highly entropic users. Low entropic users have a rather monotonous life characterized by few routines. A property of routines is their regularity, which greatly facilitates their prediction. Highly entropic users have a rather richer life with variety, freedom from routine and tend to explore and experience new locations. Therefore, prediction of their mobility is associated with great uncertainty. The location history of highly entropic users contains fewer regular patterns, thus their mobility is less

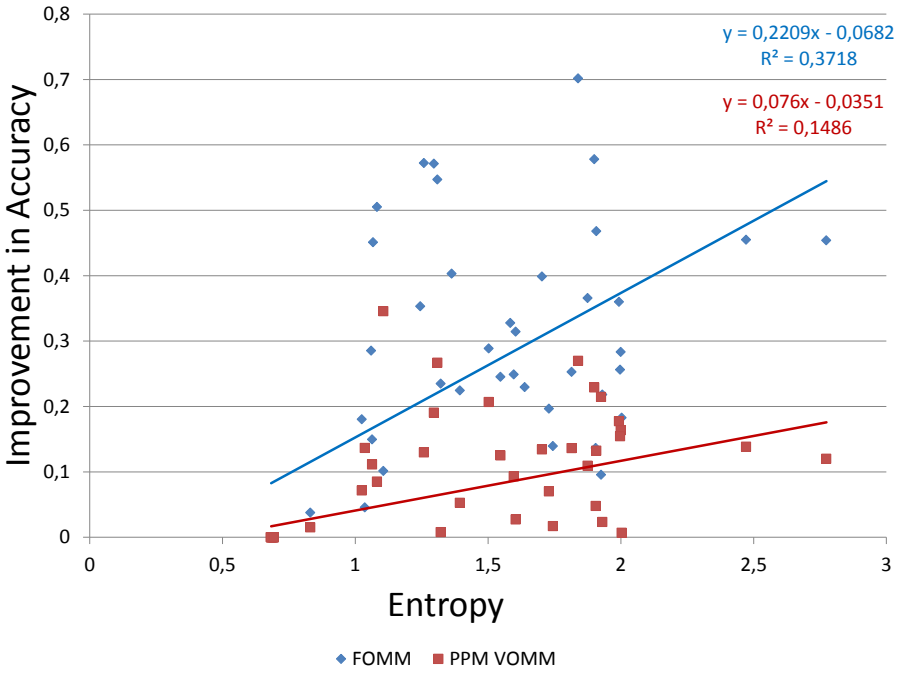

Figure 3.11: A positive correlation between the entropy and the absolute improvement in accuracy using ST PPM VOMM exists compared to both FOMM ($r = 0.61, \rho = 0.63, P(\epsilon) = 0.0$) and spatial PPM VOMM ($r = 0.39, \rho = 0.55, P(\epsilon) = 0.0$).

predictable. Although ST PPM VOMM is less expressive than HMMs and simpler in structure, we believe that the properties already mentioned compensate for the improvements in accuracy that could be achieved using highly expressive mobility models (assuming that the non-trivial training of these models converges against the true probability distribution).

Chapter 4

Correlation Between Social Network and Mobile Homophily

Homophily refers to the tendency of individuals to associate with similar individuals. Similar individuals have in common features such as interests, beliefs, thoughts, locations, emotional needs, goals, norms, etc. Mobile homophily refers to the tendency of similar individuals to be interested in the same locations, thus it involves two measurements, namely social proximity and mobile proximity. The focus of this chapter is on showing statistical dependence (if any) between social and mobile proximity. Due to cognitive, emotional, spatial and temporal limits, people cannot maintain of all their social relationships with the same intensity [Granovetter, 2005]. Each user can have subgroups of strong social relationships, in which almost everyone knows everyone else, thus the cohesion among the members of the same subgroup is considerably higher compared to a subgroup of arbitrary selected friends. The correlation between cohesive subgroups and mobile homophily is a further subject of this chapter.

Chapter summary: Section I provides a brief introduction in social networks, possibilities arising through technological achievements and the growing interest in investigating the influence of social networks in real life phenomena in the last two decades. Section II contains a formal description of social networks, their properties and interior structure. Further, the section provides an introduction of various types of locally dense regions representing cohesive subgroups, their properties, methods for their detection and a measurement for calculating group cohesion. Section III provides concepts for calculating social proximity among users. Section IV contains an introduction to the propinquity effect and methods for calculating mobile proximity. Section V presents some relevant works about the interplay between spatial properties such as locations and distance and social networks. Section VI provides an in-depth correlation analysis between social and mobile proximity based on data collected from March to July 2012 from an LBSN platform called Foursquare. The section sets particular focus on the effects of propinquity and cohesive subgroups on the mobility behavior of mobile users.

4.1 Introduction

The behavior of a mobile user cannot be explained by focusing solely on the user's individual properties defined for that individual (such as age, gender, etc.), but rather by understanding the underlying social network to which the individual belongs and by quantifying the amount of influence the social network exerts.

An individual's mobility and other behavioral aspects are impacted by their social network. Research into social networks has suffered from a lack of (real life) data, both in terms of precision and quantity. Until recently, researchers used data from questionnaires which had been answered by only a small number of respondents, artificially created simulation data, or data collected within the course of research projects ([Eagle and Pentland, 2006] and [Zheng et al., 2009a]) with a few hundred test persons to investigate the structure of communications & social networks [Onnela et al., 2007]. Such collections of data do not provide sufficient understanding of the structure and properties of large social networks with millions of users or produce much insight into the interaction paths or tie strength between the individuals in the social network.

Technological improvements and specially the increasing availability and accessibility of social media opened up new opportunities of collecting huge amount of real life data with millions of users and even more interaction/communication paths between them. Social media allows users to interact and share with each other. This has led to the formation of new online interaction and sharing sites such as Facebook [fac, 2013], Twitter [twi, 2013], etc. Further, the pervasiveness of inexpensive and precise location-acquisition technologies and their integration into mobile devices has enabled new service platforms to be built which combine both social networking and location-acquisition. Examples of location-based social networks (LBSN) are Foursquare [fou, 2013] and Plazes [pla, 2013]. Through these platforms, users can log their position into a central database and share their data with their friends. Such platforms are ideal for investigation of the spatial properties of social networks as well as the influence of social networks on an individual's mobility.

Interest in social media has been growing since the mid 90s and recently the focus has turned to LBSN. LBSN have opened up new perspectives for researchers investigating the influence two interacting users have on each other and the resulting change in their behavior. The mutual influence on the individual mobility of users who are connected via a social network is the subject of this and the following chapter. Whereas the history of an individual's movements is sufficient to predict locations that are frequently visited such as home or work, the social network of a user provide valuable information which can be used to predict locations that have been rarely visited or not yet visited by the user for the first time.

The concept of homophily states that individua tend to associate with others who have similar interests, beliefs, thoughts, hobbies and also mobility [McPherson et al., 2001]. We investigate the interdependence between mobile homophily and social proximity in order to answer questions such as: Does the social proximity between two users imply mobile proximity? Do geographical barriers or distance have an influence on mobile proximity in the Internet age? Is there a correlation between mobility proximity and group cohesion?

4.2 Social Networks

A social network is a structure consisting of a set of individuals, i.e. organizations or natural persons, and also the ties(dyadic) between those individuals. A tie represents a social relationship between two individuals such as friendship, kinship, business relationship, etc. Social Network Analysis (SNA) is the science of analyzing this structure using a set of techniques and metrics, in order to find local and global patterns in the social network.

[Moreno, 1951] developed sociometry as a qualitative method for measuring social relationships in social networks. The most important innovation of Moreno was the development of the sociogram, which is a graphical representation G of a social network. The graph $G = \{V, E\}$ contains a set of vertices V (individuals) and a set of edges E, an edge e_{ij} represents a tie between two vertices $v_i, v_j \in V$. The sociogram is a means that allows the systematic analysis of social networks using a set of metrics on the relationships (edges E) among the individuals (vertices V), the distribution of the individuals and (local) regions of high density in the graph G. The following metrics are among those fundamental to SNA and of relevance for this work:

Size $|V|$: is simply the total number of vertices in the graph.

Degree $|N(i)|$: If two vertices v_i, v_j are connected on the graph G then an edge $e_{ij} \in E$ exists and the two vertices are said to be neighbors of each other. Let $N(i)$ be the set of neighbors of vertex $v_i \in V$, the degree of the vertex v_i represents the size of its neighbors $|N(i)|$ such that $|N(i)| = |e_{ij} \in E, v_i, v_j \in V|$.

Shortest Path or Geodesic Distance $d(i, j)$: is the minimum number of edges required to connect two vertices v_i, v_j on the graph G [Cormen et al., 2001]. If two vertices are directly connected, then their distance is one. If they are connected over an intermediary vertex v_k, then the distance $d(i, j) = 2$ and so forth. The average shortest path was popularized by the small world experiment conducted by Stanley Milgram [Milgram, 1967]. The small world experiment states that the human society is a small world, where any random pair of users is on average connected by a small number of ties (hence the term six degrees of separation) [Milgram, 1967].

Density: is a property of the graph G regarding the proportion of ties $|E|$ relative to the total number of possible ties $|V| \times (|V| - 1)$ [Xu et al., 2010], the density increases if the average degree $\mu_{|N|}$ of the vertices increases. This measure quantifies how complete the graph G is. In a complete graph an edge exists between two arbitrary vertices, thus the density of a complete graph is equal to one.

4.2.1 Clustering Coefficient

[Watts and Strogatz, 1998] introduced this measure in order to determine, whether a graph G is a small world network (by name analogy with the small world phenomenon [Milgram, 1967]). Clustering coefficient is an elementary measure in social network analysis and calculates the local density of ties among the neighborhood of a vertex v. The (local) clustering coefficient determines regions of high local density and thus is used to find the degree of cliquishness in a social network.[Hanneman

and Riddle, 2011].

Real social networks are characterized by the tendency of their members to build close-knit clumps with a high density of social connections, the various knit clumps in turn are connected with few ties (weak ties). This tendency can be explained by the principle of triadic closure which states that two friends of a user are also likely to be friends of each other [Watts, 2004]. The clustering coefficient in real social networks is considerably higher compared to randomly generated social networks due to the triadic closure between its members [Newman, 2003, Milgram, 1967], therefore it is a good indicator of the existence of a real social network.

A triangle $e_{jk}|v_j, v_k \in N(i), e_{jk} \in E$ between two neighbors of the vertex v_i exists, if an edge exists between the two vertices v_j and v_k. An open triplet consists of three vertices that are connected by two edges, thus a triangle consists of three open triplets. The clustering coefficient C_i is the relationship between the total number of triangles and the total number of possible triplets in the neighborhood of a vertex v_i. The local clustering coefficient C_i for a vertex v_i quantifies to what extent the vertex v_i is likely to build a clique in its neighborhood

$$C_i = \frac{2|e_{jk}|v_j, v_k \in N(i), e_{jk} \in E|}{k_i(k_i - 1)} \tag{4.1}$$

where $k_i = |N(i)|$ is the degree of vertex v_i. The network clustering coefficient C is calculated by averaging the local clustering coefficients $C = \frac{1}{|V|}\sum_{v_i \in V} C_i$ of all vertices V of the graph G, where $|V|$ is the total number of vertices in the graph. Note that equation (4.1) calculates the clustering coefficient in an undirected graph, for directed graphs one needs to remove the factor two, because in directed graphs $e_{ij} \neq e_{ji}$ and the total number of open triplets is thus $k_i(k_i-1)$ instead of $k_i(k_i-1)/2$.

4.2.2 Cohesive Subgroups

According to modern social psychology, subgroups have different emergent features and properties, which cannot be easily deduced from the dyadic relationships of the group members. The most common properties that are agreed by many scientist are: a small size of the group so that each pair of group members can communicate directly without mediators [Homans, 2001], homogeneity of the group members with respect to common goals, norms, thoughts, beliefs, a specific communication structure and finally an emergent group-awareness [U.Brandes and Erlebach, 2004, Fischer and Wiswede, 1997].

A cohesive subgroup in social network is a tightly-knit clump with tight connections and includes the following properties ([Balasundaram. et al., 2009] and [Wasserman and Faust, 1994] as cited by [Groh, 2005]):

- **Cohesion/Density:** The density of ties in a cohesive subgroup is significantly higher compared to randomly chosen graphs.
- **Familiarity:** Most of the members have ties with every other member of the subgroup, in other words the subgroup contains few strangers.

- **Mutuality:** Most ties in the subgroup are reciprocal.
- **Compactness/Robustness:** The average shortest path within the subgroup is very short, which prevents the subgroup from being destroyed if some ties are removed.
- **Separation:** Members of the group have more ties within the group than outside.

A subgroup can be formalized as a subset $U \subseteq V$ of the vertices of a social network represented by the graph $G(V, E)$.

4.2.2.1 Cliques

U is a clique if the subgraph $G[U]$ induced by U is complete, which means for any pair of users $u_i, u_j \in U$ a tie exists $e_{ij} \in E$. A clique U is maximal if it is not a proper subset of another clique U', i.e. $U \subset U'$ is completely contained in U' and a vertex v_k exists such as the $v_k \in U'$ and $v_k \notin U$ [Luce and Perry, 1949]. A clique U is maximum if there is no other clique U' of a larger size$|U| < |U'|$ in G. Cliques are nested, because a clique of size n contains n cliques of size $n-1$, and they are robust in terms of member exclusion, because a clique without a member is still a clique ([Wasserman and Faust, 1994] as cited by [Groh, 2005]). According to this definition, a clique is an ideal cohesive subgroup with a prefect network diameter, an average shortest path and a density of one.

The completeness restriction is very strict and makes the concept of clique impractical, because it makes a clique very sensitive to missing links. In real-life scenarios, cohesive subgroups are not ideal and a few ties may not exist. Additionally, a few ties may also be missing due to errors during data collection. Clique relaxation is therefore important to make a cohesive subgroup less sensitive to missing ties. Different clique relaxation methods are based on the relaxation of one or more of the structural aspects of a clique, namely either distance, diameter or degree [Balasundaram. et al., 2009].

4.2.2.2 Distance & Diameter-Based Relaxation

N-clique: Given a graph G and a minimum distance n, the N-clique approach allows a vertex v_i to be a member of an N-clique, if the distance between v_i and any other member v_j of the N-clique $d_{(G)}(i, j) \leq n$ is less than n in the graph G. The distance restriction of a clique U is relaxed in an N-clique so that $d_{(G)}(i, j) \leq n, \forall v_i, v_j \in U)$ [Luce, 1950].

N-cliques do not have meaningful complementary definitions [Balasundaram. et al., 2009]. They tend to find wiry and long subgroups instead of tight and discrete knits [Hanneman and Riddle, 2005] because the distance restriction is with respect to the global graph G and not with respect to the induced N-clique $G([U])$ (because the neighbors $N(j)$ of each vertex $v_j \in U$ of a clique are also assigned to $G([U])$). Moreover, two vertices of an N-clique can be connected over other vertices, even without these vertices having to be members of the clique [Hanneman and Riddle, 2005].

N-clubs: The N-club approach restricts the diameter of the induced graph $G([U])$ to be less than n [Alba, 1973, Mokken, 1979].

N-clans: To overcome the drawbacks of N-cliques, the N-clan approach restricts an N-clique to being maximal and the diameter of the induced graph $G([U])$ to being less than n [Mokken, 1979], thus the N-clan approach is a minor modification of the N-clique approach.

All the previous approaches are based on relaxing the distance restriction of U, which is generally small in social networks [Milgram, 1967], thus they are less suitable for finding cohesive subgroups [Seidman and Foster, 1978] (as cited by [Balasundaram. et al., 2009]). Moreover, these approaches are neither nested, nor closed under exclusion, which makes them even less suitable for finding cohesive subgroups [Everett, 1982] (as cited by [Kosub, 2004]).

4.2.2.3 Degree-Based Relaxation

N-plexes: The N-plex approach overcomes the drawbacks of distance-based relaxation by relaxing the restriction of a clique that any two arbitrary members of the clique must be connected on the social network graph. Each member can have at most n missing ties instead of being connected to all other members of the clique [Seidman and Foster, 1978] (as cited by [Balasundaram. et al., 2009]). $U \subseteq V$ is a k-plex in $G = (V, E)$ if $|N_{G([U])}(i)| = |N(i) \cap U| > |U| - k \forall v_i \in U$ [Seidman and Foster, 1978] (as cited by [Balasundaram. et al., 2009]). K-plexes are nested and closed under exclusion and they have a diameter less than or equal to two (as cited by [Kosub, 2004]). A K-plex U is maximal if it is not a proper subset of another K-plex U' so that all vertices $v_i \in U$ are also contained in $v_i \in U'$ and the K-plex U' have at least one more vertex than the K-plex U (as cited by [Balasundaram. et al., 2009]).

4.2.2.4 Maximal Clique Detection

The maximal clique detection aims to enumerate all maximal cliques in a graph. The clique detection problem is a NP-hard problem [Makino and Uno, 2004, Wakita and Tsurumi, 2007, Tsukiyama et al., 1977, Kosub, 2004]. Enumerative algorithms contain a set of efficient algorithms that can output all possible configurations in a polynomial total time bounded by the size of both input and output. [Tsukiyama et al., 1977] has proposed an algorithm for finding all maximal cliques with a polynomial delay, i.e. the time delays between the start of the algorithm and the first output, between two consecutive outputs and between the final output and termination are polynomially bounded by the size of the input. The algorithm is based on a binary tree with n levels, where each level corresponds to a vertex $v_i \in V$ and the nodes at each level i represent a clique in $G[v_1, v_2, ..., v_i]$. The leaves of the tree represent all maximal cliques in G. Each node at level i has at most two children at level $i + 1$ as follows: if v_{i+1} is connected to all vertices of the clique U represented by the parent node at level i, then the node has exactly one child $U \cup v_{i+1}$. Otherwise it has two children: the first child is U itself, and the second child is $v_{i+1} \cup (U - \overline{N}(i+1))$, where $\overline{N}(i+1)$ contains all vertices of G not adjacent to

v_{i+1}. Both memory and processing time of the algorithm are bound to $O(|V|+|E|)$ and $o(|V|^3)$ respectively (detailed proof is represented in [Kosub, 2004]).

4.2.2.5 Measure of Cohesion

The measure of cohesion determines the quality of a cohesive subgroup, which in turn allows the identification of the more important subgroups among the set of all subgroups [Kosub, 2004]. A more cohesive subgroup has more ties inside and fewer outside the group. Given a cohesive subgroup and the adjacency matrix A, $\mathcal{C}(U)$ measures the degree of cohesion in a cohesive subgroup (Equation 4.2 [Wasserman and Faust, 1994, Chapter 7]):

$$\mathcal{C}(U) = \frac{\frac{\sum_{v_i \in U} \sum_{v_j \in U} A_{ij}}{|U|(|U|-1)}}{\frac{\sum_{v_i \in U} \sum_{v_j \notin U} A_{ij}}{|U|(|V-U|-1)}} \tag{4.2}$$

4.2.3 Centrality

Centrality: is a measure of social power [Xu et al., 2010] or importance of a vertex in the social network based on its influence on the other vertices in the graph [Hanneman and Riddle, 2011]. The importance of a vertex can be calculated based on either its degree (Degree centrality), its average distance to all other vertices in the network (Closeness centrality) or the number of all shortest paths between other vertices on which it lies (betweenness centrality) [Hanneman and Riddle, 2011, Wasserman and Faust, 1994]. Central users are important for the flow of novel information across the social network, because they have many neighbours and can bridge the gap between different social groups.

4.3 Social Proximity

The intensity of social connections between two individuals varies according to the amount of social overlap between them. Members of the same social group exhibit a higher social overlap than members of two different social groups. A social group may be a family, a circle of friends, colleagues, etc. Social proximity quantifies the social overlap between two users. The higher the social overlap, the higher their proximity in the social network. This section provides a set of well-studied social proximity measurements from the field of social networking analysis (especially in the context of collaboration networks [Newman, 2001] and link prediction [Liben-Nowell and Kleinberg, 2003]):

4.3.1 Neighborhood-Based Proximity

The following methods calculate the similarity between two users based on the social overlap between them.

Common Neighbors (CN): is the absolute overlap between the friends sets of two users and is given by $CN(u_i, u_j) = |N(i) \cap N(j)|$ where $N(i)$ is a set containing the neighbors of user u_i. The more common friends two users have, the higher their social proximity.

Adamic-Adar (AA): CN does not differentiate between the common neighbors of two users, for example the probability that two users have a friend in common depends on the number of friends this user has. A user with a high degree (a popular user with thousands of friends) is a potential common neighbor of many pairs of users ($n(n-1)/2$). Adamic & Adair therefore use a normalized version of $CN(u_i, u_j)$, where the contribution of each neighbor $u_k \in CN(u_i, u_j)$ is penalized by the inverse logarithm of their degree [Adamic and Adar, 2003]

$$AA(u_i, u_j) = \sum_{u_k \in CN(u_i, u_j)} \frac{1}{log\ N(k)} \tag{4.3}$$

A popular common friend is weighted lower, because due to his popularity it is very likely that two users have him as a common friend by chance. Whereas an exclusive common friend (a user, who is connected only to this pair of users) is weighted higher because she represents a strong evidence for a higher social proximity between the two users.

Jaccard Coefficient (Jacc): Both previous measurements consider the absolute number of social overlaps instead of overlaps in relation to the total number of friends the two users have. For example, a pair of users with a social overlap of 50 common neighbors out of the union of their friends of 1000 is less significant than three common neighbors out of four. The Jaccard coefficient sets the absolute number of common neighbors in relation to the total number of friends of both users. Jaccard coefficient calculates the similarity between two sample sets by simply dividing the size of the intersection of the two sets by the size of their union.

$$Jacc(u_i, u_j) = \frac{|N(i) \cap N(j)|}{|N(i) \cup N(j)|} \tag{4.4}$$

A pair of users, who have all their friends in common, have the highest social proximity, whereas a pair who have a lot of friends, but only a few in common, have a low social proximity.

4.3.2 Distance-Based Proximity

The following methods calculate the similarity between two users based on the collection of all paths between them

Katz: Proximity is calculated as a weighted sum over the set of all possible paths $paths_{u_i,u_j}$ between the two users [Katz, 1953]. The paths are exponentially damped by their length in order to weight short paths higher (Equation 4.5)

$$K(u_i, u_j) = \sum_{\ell=1}^{\infty} \beta^{\ell} |paths^{\ell}_{u_i,u_j}| \tag{4.5}$$

where $paths^{\ell}_{u_i,u_j}$ is the set of all paths of length ℓ and β is the damping factor (typically set to 0.05 [Katz, 1953])

Similar rank: This method is based on the intuition that two users are likely to be similar if they have ties to others who are similar. Proximity is calculated recursively according to Equation 4.6:

$$prox(u_i, u_j) = \frac{\sum_{a \in N(i)} \sum_{b \in N(j)} prox(a, b)}{|N(i)| * |N(j)|} \tag{4.6}$$

where $prox(u_i, u_i) = 1$

Random Walk Rank: The network proximity is the number of steps required during a random walk on the graph G in order to reach vertex u_j by starting at vertex u_i and moving to one of the neighbors of the current vertex at the next time step.

4.3.3 Density-Based Proximity

Degree of Cliquishness (DoC) quantifies to which extent the friends of two users build a cohesive group. Two users are close friends compared to others if they build a cohesive subgroup with their common neighbors. Let CL be a set containing the two users u_i, u_j and their common neighbors $CN(u_i, u_j)$, for each user in $u_k \in CL$ we calculate the number of triangles that contain u_k and two other users in CL and divide it by the total number of possible triangles containing user u_k (equation 4.7)

$$DoC(CL) = \sum_{u_k \in CL} \frac{|\{e_{mn} | v_n, v_m \in N_{CL}(u_k)\}|}{\frac{1}{2}|N(k)|(|N(k)| - 1)} \tag{4.7}$$

where $N(k)$ contains the neighbors of user u_k in G. Using the set of neighbours of user u_k in G in the denominator of equation 4.7 means that the impact of a common neighbor on the social proximity between two friends is higher, if the common neighbor has most of their ties inside the set CL of common neighbors than outside.

4.3.4 Cluster-Based Proximity

Clustering-Based Proximity improves the quality of the method of calculating the proximity between two users by clustering their ties and removing those neighbors that are less significant (have a smaller degree) and then running the proximity calculating method on the reduced graph G.

More measurements for calculating the network proximity are presented in [Liben-Nowell and Kleinberg, 2003].

4.4 Propinquity, Mobile Homophily, Tie Strength

Propinquity is the tendency of individuals to have ties with others who are physically close [Kadushin, 2012]. Physically close users attract each other more than

distant users, for example, two users living in the same building have a higher propinquity than two users living in different buildings [Festinger et al., 1950]. The effect of distance and physical proximity (as a cost factor) on social attractiveness for forming new friendships has been studied in [Lucille and Powell, 1975, Fischer and Wiswede, 1997], the probability of friendship increases as physical distance decreases (as cited by [Groh, 2005]).

Homophily or proximity is the extent to which an individual forms a tie to others who are similar, based on metrics such as physical distance, gender, race, age, occupation, co-locations, etc.[McPherson et al., 2001, Flynn et al., 2010]. The notion **mobile homophily** is used if the similarity between the two users is calculated based on measurements relating to geographic conditions such as distance and co-locations.

Tie Strength characterizes the intensity of a relationship between two individuals based on metrics like co-locations or the number of messages exchanged between them, etc. thus there is a strong association between tie strength, (mobile) homophily and propinquity.

Note: Propinquity does not necessarily imply mobile homophily. Propinquity states that physically close individuals influence each other more than distant individuals, whereas mobile homophily is the tendency of individuals to be interested in same locations, disregarding their physical proximity.

The measurements we used to calculate the mobile homophily between a pair of users include those introduced in [Wang et al., 2011]. The measurements can be split into two groups. The first group contains measurements based only on the calculation of the spatial overlap (asynchronous) between the trajectories of two users such as spatial co-location (Scol) and spatial cosine similarity (Scos). The second group contains measurements based on the spatial-temporal overlap (synchronous), i.e. both users were at the same location at the same time, for example, when two users are involved in the same social situation. A **social situation** occurs when at least two users arrive at a location within a defined time interval Δt [Groh et al., 2010], which guarantees an overlap between the stay times of both users at that location.

4.4.1 Measurements Within the Emphasize of Spatial Overlap

The probability that two randomly chosen users visit the same location increases when the time span between the two visits approaches infinity, but, friends usually, follow one another to a location within a considerably shorter time frame. For example, a user u_i can visit an interesting location and recommend it to their friends $N(i)$, the probability that a friend $u_j \in N(i)$ visits the same location under the influence of the visit of user $u - i$ is higher the shorter the time frame between the two visits. We use a time frame of one week for some of the following measurements. The following measurements are based on quantifying the amount of spatial overlap between the trajectories of two users using a time frame of one week.

Spatial Co-location Count (Col): a **spatial co-location** is a location visited by two users, but not necessarily at the same time. **Spatial Co-location Count (Col)** is simply counts the cases in which a visit by one of the users to a location is

followed by a visit of the other user to the same location within one week.

$$Col(u_i, u_j) = \Sigma_l \Sigma_{s_l^{(u_i)} \in H_{u_i}} \Sigma_{s_l^{(u_j)} \in H_{u_j}} \Theta(W - |T_{s_l^{(u_i)}} - T_{s_l^{(u_j)}}|) \tag{4.8}$$

H_{u_i} is the set of visits of user u_i and Θ is the Heaviside step function for two visits $s_l^{(u_i)}$ and $s_l^{(u_j)}$ of both users u_i and u_j to the same location l within a time frame of one week W [Wang et al., 2011].

Spatial Co-location Rate (SCol): The probability that two users u_i and u_j visit the same location within a period of one week:

$$SCol(u_i, u_j) = \sum_{l_k \in L} p^{(u_i)}(l_k, t) * p^{(u_j)}(l_k, t) \tag{4.9}$$

where t is set to one week and the probability mass $p^{(u_i)}(l_k, t)$ is calculated by counting the cases where user u_i follows user u_j to location l_k within one week and dividing it by the total number of visits of user u_i to location l_k. Note: This measure assumes that both users visit the same location independently [Wang et al., 2011]. Equation (4.9) assumes independency between the probability of two users to visit the same location. The value of $SCol(u_i, u_j)$ must show no correlation to the social proximity between users u_i and u_j, otherwise the assumption is rejected, meaning that the movements of a user u_i is dependent on the movement of a friend u_j.

Spatial Cosine Similarity (SCos): This measurement calculates the degree of spatial overlap between the trajectories of two users, i.e. how often the two users are co-located, disregarding the times of the visits. The trajectories of a user can be represented as a multi-dimensional vector of the number of visits to each of the locations. The similarity between two vectors can be defined as the cosine function of the angle between the two lines as in equation 4.10.

$$SCos(u_i, u_j) = \frac{\sum_{l_k \in L} p^{(u_i)}(l_k) * p^{(u_j)}(l_k)}{\|p^{(u_i)}\| * \|p^{(u_j)}\|} \tag{4.10}$$

where L is a set containing all the locations that both users u_i and u_j have in common, $p^{(u_i)}(l_k)$ is the probability of user u_i being at location l_k [Wang et al., 2011].

4.4.2 Measurements Based on Spatial-Temporal Overlap

These measurements guarantee that the two users are present during the same time at the same location, meaning that the visits of both users occur synchronously. Social situations take place within small time and location scales, which makes them important for "understanding the fundamental properties of relationships between the participating individuals" [Lehmann, 2010]. A social situation can be detected using the interaction geometry (relative body orientation and interpersonal distance) [Groh et al., 2010]. Due to the lack of information we are not able to detect social situations based on interaction geometry. Nevertheless, two users who visit the same

location within a small time frame (for example one hour), underly the same social-psychological driving forces even without observable interaction between them, and thus are most probably involved in one and the same social situation. Therefore, we assume that two users are involved in the same social situation, if they visit the same location within a short time frame Δt.

Social Situation ($\mathfrak{s}$) Rate: This measurement quantifies the degree of spatial-temporal overlap between the trajectories of a pair of users, i.e. how often a pair of users were involved in common social situations. $\mathfrak{s}(u_i, u_j)$ is mass calculated by counting the number cases where two users u_i and u_j visit the same location within a time frame Δt, normalized by the total number of times when both users were observed within this time frame Δt. Setting the temporal granularity to one hour means that the two users meet at the same location within one hour as in equation 4.11:

$$\mathfrak{s}(u_i, u_j) = \frac{\Sigma_l \Sigma_{s_l^{(u_i)} \in H_{u_i}} \Sigma_{s_l^{(u_j)} \in H_{u_j}} \Theta(\Delta t - |T_{s_l^{(u_i)}} - T_{s_l^{(u_j)}}|)}{\Sigma_{s^{(u_i)} \in H_{u_i}} \Sigma_{s^{(u_j)} \in H_{u_j}} \Theta(\Delta t - |T_{s^{(u_i)}} - T_{s^{(u_j)}}|)} \tag{4.11}$$

H_{u_i} is the set of visits of user u_i, $s^{(u_i)}$ is any visit of user u_i, Θ is the Heaviside step function for two simultaneous visits $s_l^{(u_i)}$ and $s_l^{(u_j)}$ of both users u_i and u_j to occur within the time frame Δt at the same location l [Wang et al., 2011]. Two users can be at the same location, even if only one of them makes a check-in, therefore, in order to reduce the impact of missing data, the denominator of equation 4.11 counts the number of times both users were observed within a time frame of one hour, regardless of their location.

4.4.3 Weighting Factors

The following weighting factors are used to calculate weighted versions of the above measurements:

Location density: Due to a lack of other choices in the neighborhood in suburban or rural areas, two users may visit a location by chance whereas in urban areas, such as Manhattan, two users will visit a location as the result of social influence because there is a high number of other locations to chose from to visit. The density of a location is an indication of the number of other locations in the neighborhood. We made use of the GPS coordinates provided by Foursquare to calculate the density of each location. The density of a location is simply the number of neighboring locations within a radius of one kilometer. The density is regarded as a positive weight, meaning the higher the density the higher the likelihood of visiting the location under social influence and not by chance.

Distance From Home Location: following common sense, the probability that two users are co-located by chance is higher if the two users live within a short distance of each other. Each visit of the user can be weighted by multiplying it with the logarithm of the distance between the homes of the two users. Because the home locations of the users are unknown, they are learned using the GPS coordinates of the locations provided by Foursquare. Regions of high density of check-ins are calculated

for each user using DBScan clustering. The center of the region with the highest occurrence of check-ins represents the home location of the user. The center of each region is calculated by simply averaging the coordinates of the check-in locations within the region. Many studies have confirmed that people tend to visit locations nearby their homes, therefore it is reasonable to define the region with the highest occurrence of check-ins, as the home location of a user [Volkovich et al., 2012, Noulas et al., 2012, Noulas et al., 2011a, Scellato et al., 2011b]. As soon as home locations have been determined, the physical distance between two users can be calculated.

Location Population $\rho(l_k)$: People can meet by chance at busy locations which are accessible for everybody (public locations such as subways or sport stadiums). Whereas they tend to meet close friends at less busy locations such as their private domicile or a small restaurant/cafeteria nearby their home (with a small number of visitors). Let $\rho(l_k)$ represent the population visiting the location l_k. Each visit to a location l_k is inversely proportional to the log of the population size $|\rho(l_k)|$ [Wang et al., 2011].

Location Entropy: The Shannon entropy quantifies the expected average amount of information with respect to the distribution $p(x)$ of a random variable x [Bishop, 2007, page 49] given by equation (4.12).

$$H(X) = -\sum_{x_i \in X} p(x_i) \ \ln p(x_i) \tag{4.12}$$

where $p(x_i)$ is the probability of the occurrence of x_i. We use natural logarithms for calculating Shannon entropy, but logarithm of other bases can also be used.

[Russell and Norvig, 2010, Page 703] describes the Shannon entropy as the measure of uncertainty in the distribution of a random variable. The Shannon entropy for a location visited by many users with similar frequency (such as bars, restaurants or shopping malls) is high, because its value is distributed equally over a wide range. Therefore the uncertainty of predicting which users are visiting the location is high, thus knowing the users means a high information gain. In contrast, a low entropic location is visited by a relatively small number of users and with quite different probabilities. The probability of a check-in is inversely proportional to the log of the entropy of the visited location, which means a higher weight for check-ins to low entropic locations. The mobile homophily between two users is high if they meet at a private location such as home, instead of meeting at public locations such as a bar or restaurant.

4.5 Related Work

Physical distance plays a key role in many investigations regarding the interdependence between mobility and social networks. The probability of friendship as a function of distance $P(d)$ has been investigated in [Lambiotte et al., 2008, Nowell et al., 2005, Backstrom et al., 2010, Scellato et al., 2011a, Volkovich et al., 2012, Scellato et al., 2011b]. The influence of mobility on the formation of new social ties has been investigated by [Wang et al., 2011, Cho et al., 2011, Scellato et al., 2011a, Lee

et al., 2011, Liben-Nowell and Kleinberg, 2003]. The influence of social networks on predicting individual mobility has been investigated by [Sadilek et al., 2012, Eagle and Pentland, 2009, Cho et al., 2011, Hackney and Axhausen, 2006, Tang et al., 2009]).

[Volkovich et al., 2012] has investigated the relationship between spatial distance and social interactions. It has been shown that the social overlap between pairs of friends increases as their geographical distance decreases. The authors found that the geographic distance between two users is inversely correlated to the size of their common neighbors. Users with a high social overlap tend to live within a small geographical distance. The average shortest path among users living within a geographical distance of less than 60-80 km is rather short. The amount of social overlap exhibits a rapid reduction and the average shortest path increases considerably if the geographical distance exceeds 60-80 km. Therefore the authors suggest dividing social ties into inter and intra city connections with a separation distance of between 50-100 km.

[Kaltenbrunner et al., 2012] found that online interaction between users is not affected by their geographical distance. Online social networks are formed on a global scale without restriction to a town or a country, therefore interaction in online communities takes place irrespective of geographical conditions. The influence of technological factors on spatial proximity has been investigated by [Goldenberg and Levy, 2009] by analyzing the spatial dissemination of new baby names. In contrast to the opinion that the internet has reduced the influence of geographical distance on social interactions, the authors found that although technological advances have caused an increase in the overall amount of communication, most of this communication occurs between spatially close users.

The prediction of user addresses based on the addresses of their friends has been investigated in [Backstrom et al., 2010] using a maximum likelihood estimator. The probability of friendship vs. distance of Facebook users was found to follow a power law. The influence of physical distance on the formation of new links between unconnected users and on tie strength between already connected users has been investigated by [Wang et al., 2011, Cho et al., 2011, Scellato et al., 2011b]. The tie strength as a function of distance is investigated by [Scellato et al., 2011a] and suggests that expensive and distant connections appear mainly between the members of a social network who are important to each other. The relationship between local network topology and tie strength on global information diffusion, as well as the stability of social networks when links are removed are the subjects of a study in [Onnela et al., 2007].

A strong correlation between social networks and mobility has been shown in [Gonzalez et al., 2006]. The authors were able to reproduce the social network based on an analysis of the motion of its members. The degree distribution, clustering coefficient, and the average shortest path of the network were reproduced.

Prediction of new friendships among "place friends" has been investigated in [Scellato et al., 2011c]. "Place friends" are two users who have visited at least one common location, but are not yet socially connected on the LSBN network. It has been shown that 30% of newly-formed friendships are between "place friends". This result confirms the causal relationship between spatial similarity and friendship formation.

The correlation between co-locations and social relationships, as well as the calculation of the probability of friendship based on measurements like diversity/specificity (shopping mall vs. home) of co-locations and the temporal regularity of visits (such as hours of day and days of week) has been investigated in [Cranshaw et al., 2010]. The experimental results show that users with irregular mobility behavior, i.e. users who visit locations with a high diversity (such as shopping malls), tend to have a lot of friends. Whereas users having co-locations at more specific locations (such as home locations) tend to have stronger social relationships.

[Crandall et al., 2010] has investigated the correlation between co-locations and social tie. They found out that the probability of friendship increases if two users have more co-locations within a smaller time frame. A similar study by [Dong et al., 2011] shows that the number of unique co-locations between friends increases as time goes on. [Zheng et al., 2011] has shown that a strong correlation between social proximity and mobility exists and that the tie strength between users having similar mobilities is higher. Based on the number of common dishes ordered in restaurants among friends, [Lee et al., 2011] has investigated the effects of social influence, co-present and virtual influence on individual behavior. The experimental results show the importance of social networks, besides time and the current context of the user, in explaining an individual's behavior.

[Noulas et al., 2012] has studied the mobility pattern of mobile users living in urban areas based on a Foursquare data set from several metropolitan cities. In their study, the authors defined a new distance measurement between a start and destination location as rank-distance. Rank-distance differs from physical distance, because it takes the spatial distribution of locations of interest into account. They follow the intuition that the interest of mobile users in distant locations decreases if the number of intervening locations increases. Mobile users are rather more interested in nearby locations with a shorter travelling distance. Their experimental results confirm this intuition as the probability of moving from the current location to a possible destination is inversely proportional to the number of intervening locations. This result implies that the travel distance is inversely correlated with location density of the area. Users living in denser areas (such as metropolitan areas) have shorter travel distances than users living in less dense areas.

Taking the benefits produced by categorizing places to make similarity comparisons between users and locations, [Noulas et al., 2011b] found that the likelihood of mutual influence is higher between users visiting similar locations. The comparison was carried out by representing each user with a vector of location categories. For each user, the weight of each category is calculated based on the past visits of the user to locations in the respective category.

[Wang et al., 2011] studied the correlation between human mobility and social ties. Their work was based on large-scale mobile phone data collected by a mobile phone provider containing 90 000 000 communication records of more than 6 000 000 users, and over 10 000 distinct locations covering a radius of more than 1 000 kilometers. Using the answers to questions such as, "How connected are two users in a social network?", "How similar are the movements of two users" and "How intense is the interaction between a pair of users" the study focused on a correlation analysis on mobility and social ties. The analysis provides a good foundation for the design of

a complete correlation analysis, e.g. taking into account the fact that real friends usually meet at certain times of the day and at certain unique locations (unlike random people with no social ties who commute daily in a subway station). The following correlation analysis was inspired to a great extent by this study [Wang et al., 2011].

4.6 Correlation Between Social Proximity and Mobile Homophily

Correlation analysis measures the amount of statistical dependence between two random variables. This section describes a deep correlation analysis of mobile homophily and network proximity measurements based on a dataset collected from a location-based social network platform called Foursquare.

4.6.1 Foursquare, an Online Location-Based Social Network

Geosocial networking or location-based social networking is a type of social networking where functionalities related to geographic locations are combined with social networking functionalities in order to offer services of higher quality. Foursquare provides a mobile application and a web platform for location-based social networking [Fou, 2012b]. Registered users can check-in at venues using the web page of Foursquare or a device-specific application (such as a smart phone app) provided by Foursquare. A check-in can be performed by selecting a venue (location) from a list of venues that the application locates nearby (therefore the use of a Foursquare data set requires no preliminary clustering for finding significant locations). The venues in Foursquare are very fine grained so that you can distinguish between two floors of the same building or two rooms of the same flat. The application maintains a social network that allows users to connect and communicate with their friends. Foursquare offers developers an API for integrating the Foursquare functions in their applications [Fou, 2012a]. We used the Foursquare API to extract a list of venues in San Francisco containing 46 853 venues together with their details and check-in statistics. For these venues we extracted all check-ins made by Foursquare users between March 23, 2012 and July 23, 2012. Each check-in is a triplet $< u, l, t >$, consisting of the user's id, the check-in location id and the date/time of the check-in. We extracted $\approx$ 2 million check-ins during the four month period.

4.6.1.1 The Social Network

The Foursquare dataset contained 141 750 users, who have generated at least one check-in. The dataset also contained the friends and the friends of friends (FoF) for each user. Table 4.1 shows the user statistics of the dataset. We refer to users who have generated at least 50 check-ins as active users. Users who have generated fewer than 50 check-ins may be tourists who visit San Francisco for a short time or people from outside whose regular life does not take place in San Francisco. The

type	All	active
# Users (U):	141 750	9 173
# ties of Users (t_U):	5 327 041	618 970
Av. degree ($\frac{U}{t_U}$):	37.58	67.59
# Users + Friends (UF):	1 747 783	261 780
# ties of User + Friends (t_{UF}):	74 585 447	25 293 730
Av. degree Users + Friends ($\frac{UF}{t_{Uf}}$):	42.67	96.62
# Users + Friends + FoF:	7 954 935	1 155 324

Table 4.1: Social Network statistics.

amount of location data generated by those users is not sufficient for making a reliable correlation analysis, therefore they are excluded from the set of active users.

Table 4.1 shows, an average degree of 42.67 among the users with at least one check-in, their friends and their friends of friends. The average degree increases to 96.62 considering only the users with at least 50 check-ins, their friends and their friends friends, because users with more check-ins rather are active since a longer period of time on Foursquare, compared to new users with few check-ins. We consider Foursquare as a hybrid social network, because Foursquare, unlike pure online social network platforms such as Facebook, motivates their users to share real life movements with their friends, thus we believe that Foursquare users spend less time on the online platform. This assertion becomes more clear when comparing the average degree in our Foursquare data set to the average degree of offline and online social networks. The average number of social links of the users is significantly higher when compared to most of the (offline) social networks listed in [Newman, 2003], and lower compared to pure online social network platforms like Facebook (the average number of degree of Facebook is currently found to be 190 [Ana, 2013]). Further, Foursquare is newer than Facebook, which is possibly another reason for the lower average degree in Foursquare compared to Facebook.

The average degree between active users and their friends is 67.59 (set U), which is less than the average degree of 96.62 between the active users, their friends and their friends of friends (set UF). The first set U is considerably smaller than set UF and contains mainly active users, who have an explorative real world life and spend a considerable amount of their times exploring the world rather than to sit in front of their PCs. Therefore, set the average degree among the users in set U is considerably smaller than the average degree among the user in set UF.

We construct two social networks from the dataset. The first social network contains the users who have generated at least one check-in and the second social network contains only the active users. We add to each of the two social networks the friends and the friends-of-friends (FoF) of the users. The social network induced by the active users has a significantly higher average degree of 67.59 compared to the average degree 37.58 of the social network induced by all users, which means that the active users build a significantly denser close-knit social network, possibly because the set of active users rather contains users who are active since a longer period of time on Foursquare.

As mentioned earlier, the clustering coefficient is a good indicator of the existence of a real social network among a set of users and their relationships [Watts, 2004]. The average clustering coefficients (equation 4.1) for the social network graphs induced by the all users and the active users were found to be 0.104 and 0.1438 respectively (table 4.2). In order to assert that the Foursquare data set represents a real social network, it must be compared to the clustering coefficient of a randomly generated social network with the same number of vertices and the same average degree. The average degree of the social network induced by all users is 42.67, and the probability of an edge existing was found to be 0.00044 [Jesdabodi, 2012]. Based on these values, a random graph was constructed in [Jesdabodi, 2012] using the Poisson random graph method presented in [Newman, 2003]. The clustering coefficient of the random graph was found to be 0.0002 on average, which is considerably lower than the clustering coefficient of the social network induced by all users 0.104 (around factor 520), and even lower when compared to the clustering coefficient for the social network induced by the active users 0.1438 (around factor 719). Therefore the conclusion is valid that both social networks induced from the Foursquare data set represent two valid social networks. Further, it is worth mentioning, that the clustering coefficient of the social network induced by the active users is around %38 higher than the clustering coefficient of the social network induced by all users. This result is in accordance with the average degree of both graphs, stating that the active users together represent a denser close-knit group.

type	**All**	**Active**
Mean average path length:	4.152 ± 0.58	3.8 ± 0.57
Clustering Coefficient:	0.104	0.1438

Table 4.2: The mean average path length is calculated using the Breadth First Search (BFS) algorithm, which has a space complexity of $O(|V| + |E|)$ and time complexity of $O(|V| + |E|)$ [Cormen et al., 2009].

The mean average path length, as stated earlier, is an indicator of whether a social network is a small world, generally, a small world is given if the mean average path length is smaller than $O(log(n))$, where n is the number of users in the social network [Newman, 2003]. The calculation of the average shortest path is a very resource-demanding task, in terms of both memory and time. Therefore we did the calculation based on 100 000 randomly chosen pairs of users from each of the social networks. The average shortest path for the entire users was found to be 4.15 ± 0.58 whereas the active users have an average shortest path of 3.80 ± 0.57, which is indeed smaller than $O(log(n)) = 6.24$ for the entire users and $O(log(n)) = 5.41$ for the active users. The mean average path lengths in both social networks is also in line with Stanley Milgram's small world phenomenon [Milgram, 1967, Travers and Milgram, 1969]. It is worth mentioning at this point that the mean average path length among the active users is shorter than the mean average path length among all users. The active users are probably rather users from San Francisco, because it is unlikely that a user can generate 50 check-ins within a 4 month period in a foreign city. Although an active user can have friends and friends-of-friends from elsewhere than San Francisco, nevertheless the mean average path length among them is smaller. The smaller mean average path length suggests that users from

the same city build a small world, which is a first indication of the existence of the propinquity effect.

4.6.1.2 The Check-In Behavior of Foursquare Users

Within the four month period (122 days) of data collection, a total number in excess of 1 983 772 million check-ins was generated by 141 750 users at 30 630 locations in San Francisco. The active users were responsible for the generation of 1.14 million check-ins. The active users built just 6.46% of the total number of users, which means that 6.46% of the users are responsible for the generation of 57.62% of the total check-ins. Table 4.3 shows the check-in statistics from two perspectives, namely from the perspective of the users who generated the check-ins, and from the perspective of the check-in locations (The ± numbers represent the average standard deviation σ).

type	**All**	**active**
# Users:	141 750	9 173
# Locations:	30 630	26 780
# Check-in count:	1 983 772	1 164 085
Av. Check-ins per user and day	0.12	1.04
Av. Check-ins per location:	64.77 ± 436.17	43.47 ± 139.72
Av. Check-ins per user:	13.99 ± 37.57	126.90 ± 84.28
Av. Locations per user:	8.71 ± 17.45	62.35 ± 31.99
Av. Users per Location	40.33 ± 282.80	21.36 ± 71.11
Av. Check-ins per user and location	1.61 ± 3.35	2.04 ± 4.74
Av. Degree of repetition:	0.61 ± 3.35	1.04 ± 4.74
Av. User entropy	1.15 ± 0.99	3.48 ± 0.71
Av. Location entropy	1.73 ± 1.60	1.66 ± 1.57

Table 4.3: Check-in and location Statistics.)

We calculate user entropy for a user u_i, who visits a set of locations L according to equation 4.13:

$$H^{(u_i)}(L) = -\sum_{l_k \in L} p^{(u_i)}(l_k)\ \ln p^{(u_i)}(l_k) \tag{4.13}$$

where $p^{(u_i)}(l_k)$ is the probability that user u_i visits location l_k.

We calculate location entropy of a location l_i visited by a set of users U according to equation 4.14:

$$H^{(l_i)}(U) = -\sum_{u_k \in U} p^{(l_i)}(u_k)\ \ln p^{(l_i)}(u_k) \tag{4.14}$$

where $p^{(l_i)}(u_k)$ is the probability that location l_i is visited by user u_k.

The performance of the mobility models in Chapters 2 and 3 were evaluated using only users who have visited at least 3 locations, and each of these locations were visited at least 5 times by them. In other words, only significant locations of the users (such as home, work, etc.) were used. Locations with a frequency of visit of less than 5, represent less significant locations of a user, which are not considered until now due to the lack of evidence that would allow their prediction. Locations with a frequency of visit per user of less than 5 can be assumed be less significant for the user, because these locations are more publicly accessible and less specific for single users. Shopping malls, hospitals, etc. are examples of non-significant locations.

Foursquare categorizes its locations into many different categories, where each location can be in zero categories or more categories. In Figure 4.1 we plotted the categories and the number of check-ins at locations falling into each category.

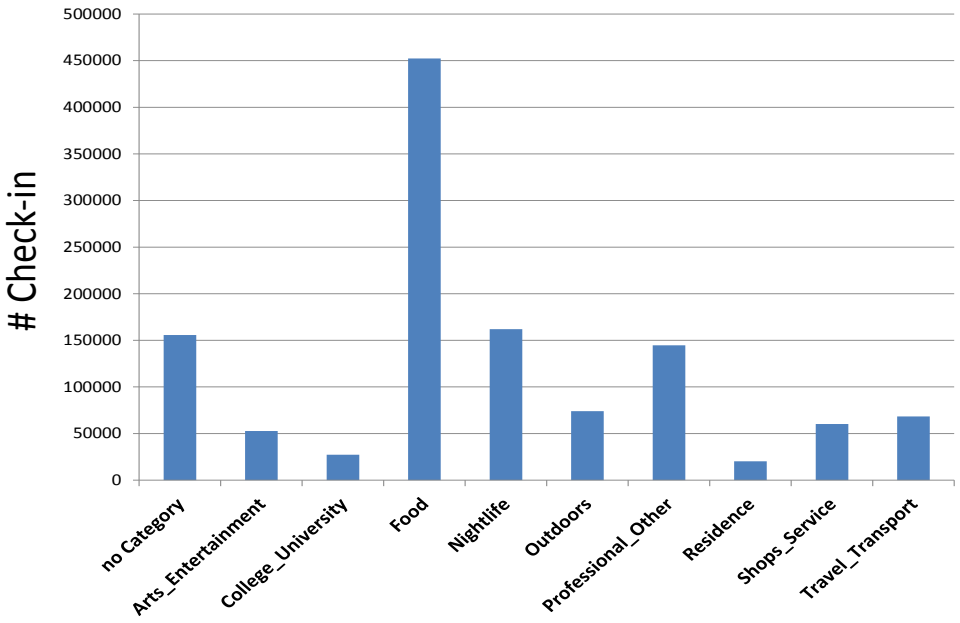

Figure 4.1: The distribution of check-ins over different categories of locations. Foursquare provides many more categories of locations, for simplicity reasons we have merged the categories into nine different category groups.

Indeed most of the categories shown in Figure 4.1 contain publicly available locations with a low frequency of visit per user. The check-in statistics show that as many as 52% of the total check-ins, and 38% of the check-ins of the active users are non-recurring check-ins. Around 60% of the check-ins were generated by users who visit the location with a frequency of less than five times. Furthermore, 33.63% of the locations have an average check-in per user of one.

Figure 4.2 shows the average history size per user for the locations. Only about 3.4% of the locations have an average history size per user greater than 10 and only 26 locations have a value greater than 50. The maximum history size per user is 102 for the only location having a value greater than 100. The Shannon Entropy for locations and users are high, i.e. many locations are visited by many

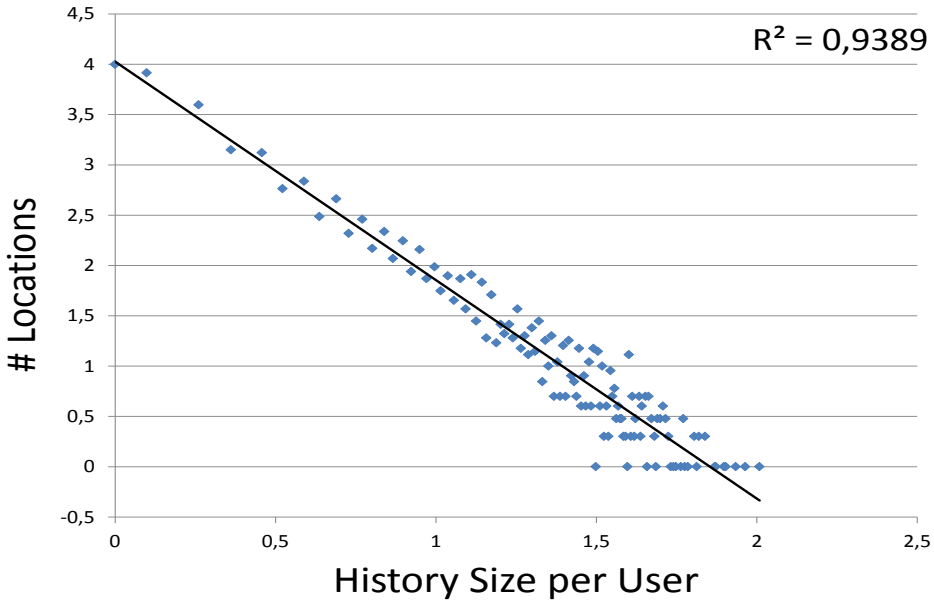

Figure 4.2: A Log-log plot representing the relationship between the average history size per user(x-axis) and the number of locations (y-axis).

users with similar probabilities, and many users also visit many locations with similar probabilities. For example, assuming that a user visits every location with the same probability, an entropy value of 3.48 means that the user can be found in one of $e^{3.48} = 32.46$ locations.

These statistics show that the Foursquare dataset contains mainly check-ins to public locations that are accessible for everyone with a low frequency of visit per user. Each active user makes on average ≈ 1 check-in per day and $\approx 2.04 \pm 4.74$ check-ins per location and visits 62.35 ± 31.99 locations, which explains the high average user entropy of 3.48 ± 0.71 and location entropy of 1.66 ± 1.57.

4.6.2 Empirical Results of the Correlation Analysis

Correlation refers to any linear statistical relationship between two random variables X and Y and is an indication of statistical dependency between X and Y. Pursuing the idea of homophily, the following correlation analysis should provide evidence about the existence of any linear statistical dependence between two visits of two users u_i and u_j to a location and their social relationship. Although correlation analysis disregards the order of the visits of the two users (whose visit follows the other), which is important for showing causation effect, it is still a good indicator of the existence of predictive dependence between the social relationship and the mobility of the two users.

Table 4.4 presents the first evidence for the existence of statistical dependence between social ties and mobility. Two neighbors share on average 4.29 common locations and 5.74 social situations, while random pairs share 1.61 common loca-

tions and just 0.14 social situations. This preliminary finding suggests that social relationships may affect the mobility behavior of individuals, because the average common check-in among friends is 2.5 times greater than the corresponding value for a random pair of users.

Type	ϕ **Common locations**	ϕ **Social situations**
Friends:	4.29	5.74
Random pairs	1.61	0.14

Table 4.4: Friends have on average 2.5 times more common locations than random pairs of users and are involved in about ≈ 40 times more social situations within 1 hour.

The following correlation analysis is based on the social proximity and mobile homophily measures presented in sections 4.3 and 4.4. For the correlation analysis we always used 100 000 randomly sampled pairs of users chosen from the whole population using a simple random sampling method (sampling a user randomly from the entire population without replacement, so that each user can be selected with equal probability during the entire sampling period). The chosen pairs are not necessarily connected through the social network, which is important in quantifying the actual correlation. The hypothesis is that mobile homophily correlates with network proximity, i.e. a linear statistical dependence exists between them. We define the following equivalences for certain intervals of the correlation coefficient: ≥ 0.7 corresponds to very strong correlation, $[0.4, 0.7]$ corresponds to a strong correlation, $[0.1, 0.4]$ corresponds to a moderate correlation, ≤ 0.1 corresponds to weak or non correlation. The social proximity between two users is calculated using three neighborhood (CN, AA and Jacc) and one density-based (DoC proximity measurements (Section 4.4). Mobile homophily is calculated using all the measurements in section 4.4. Following the intuition that people tend to meet close friends during their free time, an additional measurement, extra-role, is used to refer to those co-locations which occur during weekends and evening hours.

Table 4.5 shows the results of the correlation analysis. Although a correlation is noticeable between DoC and all mobile homophily measurements, the remaining measurements unfortunately show no correlation. The results of the correlation analysis are disappointing at first glance, but the low correlation may be due to measurement artifacts, because the social network is available on a global scale and not limited to San Francisco, whereas the location data is indeed limited to San Francisco. Two users might be very similar on the social network side, but unfortunately due to the lack of location data, no reliable correlation can be determined between mobile homophily and social proximity.

4.6.2.1 The Propinquity Effect

To reduce the effects of measurement artifacts and to verify the propinquity effect, we have repeated the correlation analysis and restricted the social network to users from San Francisco. Foursquare provides the home city for each user so that the portion of users from the same city can be extracted. Each user has on average a degree of 67, but only 20 from his home city, which is only $\approx 30\%$ of the total number

	CN	AA	DoC	Jacc
Scos:	-0.056	0.008	0.16	0.004
SCos-User:	-0.058	0.008	0.15	-0.017
SCos-Dens:	-0.052	0.013	0.155	0.013
SCos-Dist:	-0.037	0.014	0.155	0.024
SCos-Entr:	-0.048	0.008	0.155	0.004
Scol:	-0.171	0.021	0.131	0.016
Col:	-0.207	0.023	0.131	-0.007
s:	-0.066	0.043	0.112	0.021
s-Dens:	-0.027	0.048	0.105	0.022
s-Dist:	-0.01	0.07	0.084	0.014
s-User:	-0.029	0.043	0.109	0.014
s-Entr:	-0.029	0.048	0.108	0.019
s-Extra Role:	-0.15	0.009	0.118	0.034

Table 4.5: Pearson's correlation coefficient r between mobile homophily and network proximity for 100 000 randomly chosen pairs of users (setting $\Delta t = 1$ hour for spatial-temporal overlap).

of friends. Figure 4.3 shows the number of friends from the home city compared to the degree of the users.

xxx

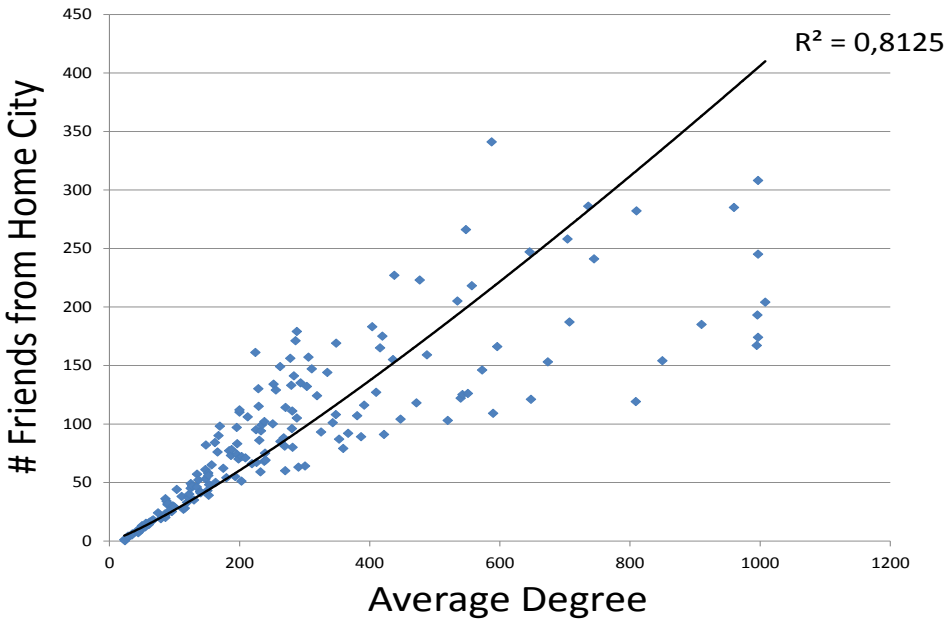

Figure 4.3: A plot representing the average number of friends from the home city compared to the degree of the users (online friends).

In accordance with Tobler's first law of geography, "Everything is related to everything else, but near things are more related than distant things" [Tobler,

1970] and in accordance with the propinquity effect, the social network formed by the active users and their friends and FoF from San Francisco (home city) has a higher average degree 67 compared to total users 37. The clustering coefficient and the average shortest path for the graph G_{HC} induced by the new social network dramatically increases when compared to the graphs induced by the total users G and the graph by the active users G_{AU}. The clustering coefficient of G_{HC} is found to be $C_{G_{HC}} \approx 0.38$, which is significantly higher than the clustering coefficients $C_G \approx 0.104$ and $C_{G_{AU}} \approx 0.1438$. The average shortest path increases from 4.152 ± 0.58 for G to 3.6745 ± 0.55 for G_{HC}. A two-sided unpaired t-test confirms the significance of the change in the average shortest path $p(\epsilon) = 1.09 * 10^{-294}$).

Type	**Clustering Coefficient**	ϕ **Shortest path:**
All:	0.104	4.152 ± 0.58
Active users:	0.1438	3.8 ± 0.57
Active home city	0.38118	3.6745 ± 0.55

Table 4.6: A comparison between the results of both clustering coefficients and average shortest path for three different graphs induced by all users, active users and users from San Francisco respectively.

A new correlation analysis using random pairs of users from G_{HC} substantiates the above findings. Table 4.7 shows how the correlation between mobile homophily and network proximity significantly increases. The significance of the changes in CN, AA, Jacc and DoC is confirmed by 3 two-sided unpaired t-tests (with $p(\epsilon)$ values $0.000601, 0.0122\&0.000124$ respectively).

	CN	AA	DoC	Jacc
Scos:	0.417	0.225	0.168	0.315
SCos-User:	0.24	0.185	0.184	0.327
SCos-Dens:	0.418	0.221	0.169	0.32
SCos-Dist:	0.389	0.154	0.153	0.273
SCos-Entr:	0.267	0.192	0.175	0.313
Scol:	0.141	0.101	0.188	0.25
Col:	0.309	0.122	0.188	0.234
s:	0.326	0.242	0.151	0.25
s-Dens:	0.309	0.281	0.16	0.341
s-Dist:	0.342	0.272	0.169	0.353
s-User:	0.215	0.231	0.178	0.349
s-Entr:	0.28	0.261	0.176	0.35
s-Extra Role:	0.052	0.163	0.145	0.351

Table 4.7: Pearson's correlation coefficient r between mobile homophily and network proximity for 100,000 randomly chosen pairs of users from G_{HC} (setting $\Delta t = 1$ hour for spatial-temporal overlap).

The correlation analysis between social and mobility proximity for users from the same home city emphasizes the propinquity effect that physically close friends influence each other more than distant users. Further, the result shows a statistical

dependency between the movements of friends, which is promises success for improving the accuracy of individual mobility prediction considering the movements of friends. Furthermore, the weighting factors distance and density increase the correlation coefficients, which means that people are willing to cover greater distances, more likely in order to meet important friends. In addition, a meeting in a location in a region dense with interesting locations, is more conscious than random in comparison to a meeting at a location in a sparsely populated region in a rural area.

4.6.2.2 The Effect of Cohesive Subgroups

As in real-life, most of the users in online communities interact mainly with a small group of their acquaintances ([Wilson et al., 2009, Jiang et al., 2010] as cited by [Kaltenbrunner et al., 2012]). These acquaintances are usually close friends with a high tie strength, who together form with a higher probability a cohesive subgroup. This section describes an investigation into the correlation between group cohesion and mobile homophily. From the behavioral-perspective, members of a cohesive subgroup share information and have homogeneity of interests and beliefs[Balasundaram. et al., 2009], therefore they exhibit a high similarity in their behavior including mobile homophily.

The method described in section 4.2.2.4 detects all maximal cliques to which a user belongs. For each user u_i, let $G[u_i] = (N(i) \cup u_i, E_{u_i})$ be the graph induced by the user, their friends and the edges between them E_{u_i}. The graph $G[u_i]$ is considerably smaller $G[u_i] \ll G$ than the total network graph G. Every maximal clique U in $G[u_i]$ is also a maximal clique in G, otherwise it would mean that G contains a vertex v_i, that is connected to user u_i in G and is not a member of the graph $G[u_i]$. The binary tree for the detection of all maximal cliques in $G[u_i]$ has $|N(i)|+1$ levels, where $|N(i)|$ is the degree of user u_i. We could detect 8700 maximal cliques consisting of at least three vertices in the social network graph G, with the largest detected clique consisting of 17 vertices. The 8700 maximal cliques are not complete, the dataset contains pretty sure more maximal cliques. As stated earlier, the completeness-requirement of cliques makes them impractical for the detection of real-life cohesive subgroups, therefore we slightly relax the detected cliques to 2-plexes in the graph G (in order to not significantly changing the clique properties) as follows:

For each clique U, let CU be a set containing all adjoinable vertices that have only one missing edge to all vertices of the clique U. Similarly to the detection of maximal cliques, we detect 2-plexes by constructing a binary tree based on CU and U as follows: The root of the tree is U, the tree has $|CU| + 1$ levels, each level $i > 1$ corresponds to a vertex $v_i \in CU$. Each node of the tree at level i corresponds to a 2-plex PL and has one or two children at level $i + 1$. The left child is a node corresponding to PL, the second child is $PL \cup v_{i+1}$ if and only if the vertex v_i is connected to $|PL| - 1$ vertices in PL. The tree contains one or more leaves. The first leaf on the left corresponds to U, the remaining leaves (if any) correspond to all maximal 2-plexes in G that contain at least $|U| - 1$ vertices of U and one vertex of CU. Based on this procedure, the social network contains 29591 2-plexes of size $|PL| \geq 3$, the plexes contain on average 7.79 ± 2.41 vertices. The largest 2-plex

contains 20 vertices, which is in rough agreement with the human social perception limit in [Fischer and Wiswede, 1997].

We made a new correlation analysis between mobile homophily and network proximity among members of the same 2-plexes by randomly sampling 100 000 pairs of users, where each pair is chosen from the same 2-plex. We present our results in a different style compared to the style of table 4.7, because the pairs of users in the new correlation analysis are chosen from the same 2-plexes, which means all social proximity measures are always high, independently from the values of mobile proximity measures. Table 4.8 compares the average values of all mobile homophily and social proximity measurements with the corresponding values of the previous correlation analysis between pairs of users from the same home city, let G_{HC} be the graph induced by the users from the same home city. All average values are significantly higher among members of the same 2-plexes, on average the network proximity measurements are as much as a factor of 50 and the mobile homophily measurements as much as a factor of 23 higher. The results show a significantly stronger interdependence between mobile homophily and social proximity among the members of cohesive subgroups, compared to arbitrary chosen pairs of users. The result gives us confidence that an influential effect exists between social proximity and individual mobility.

	Home city	2-plex
Scos:	0.013 ± 0.012	0.11 ± 0.101
WSCos-Dens:	0.014 ± 0.012	0.111 ± 0.101
WSCos-Dist:	0.008 ± 0.007	0.055 ± 0.051
WSCos-User:	0.008 ± 0.006	0.099 ± 0.075
WSCos-Entr:	0.01 ± 0.008	0.107 ± 0.09
Scol:	0.001 ± 0.001	0.009 ± 0.007
Col:	1.626 ± 1.27	40.006 ± 29.176
s:	0.058 ± 0.03	4.538 ± 3.284
s-Dens:	0.006 ± 0.003	0.158 ± 0.124
s-Dist:	0.003 ± 0.001	0.054 ± 0.042
s-User:	0.002 ± 0.001	0.044 ± 0.035
s-Entr:	0.004 ± 0.002	0.131 ± 0.103
s-Extra Role:	0.002 ± 0.001	0.054 ± 0.037
CN:	0.193 ± 0.103	14.842 ± 17.881
AA:	0.102 ± 0.054	7.804 ± 10.526
Jacc:	0.005 ± 0.002	0.059 ± 0.112
DoC:	0.002 ± 0.001	0.082 ± 0.081

Table 4.8: The average values for network proximity and mobile homophily measurements for 100 000 random pairs chosen from 2-plexes and 100 000 random pairs chosen from G_{HC}.

In order to investigate the interdependence between mobile homophily and group cohesion, we calculated the cohesion measure for the detected 2-plexes according to equation 4.2. Figure 4.4 shows the relationship between the size and the measure of cohesion of the 2-plexes. A negative trend exists between group cohesion and group

size, the smaller the 2-plex, the higher the cohesion. The cohesion among members of groups with a size ≥ 6 decreases marginally compared to the cohesion among members of groups with a size ≤ 6.

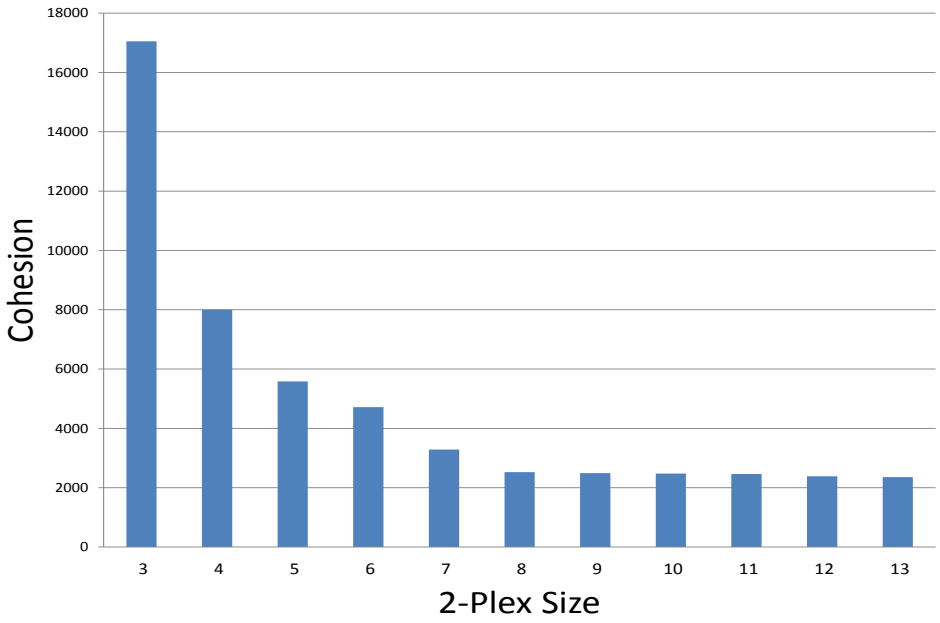

Figure 4.4: A plot representing the relationship between 2-plex size and the measure of cohesion according to equation 4.2.

Furthermore, a strong correlation exists between group cohesion and mobile homophily as shown in table 4.9. The measurements of spatial overlap (asynchronous) show a correlation coefficient varying between $r = 0.22, \rho = 0.23$ for co-locations within one week and $r = 0.51, \rho = 0.65$ for $WSCos - Entr$. The spatial-temporal (synchronous) overlap shows a higher correlation coefficient varying between $r = 0.48, \rho = 0.63$ for social situation rates ($\mathfrak{s}$)and $r = 0.66, \rho = 0.81$ for $\mathfrak{s} - Dist$, which confirms the importance of social situations for understanding the fundamental properties of social relations ([Groh et al., 2010]).

Spearman's rank correlation coefficient shows an even stronger correlation compared to Pearson's correlation coefficient due to the following two reasons. Firstly, the relationship between mobile homophily and network proximity seems to follow rather a monotonic than a linear function, and secondly because of the low sensitivity of Spearman's rank correlation coefficient to outliers. The relationship between group cohesion and mobile homophily is shown on two log-log plots on figure 4.6 and figure 4.5 for measurements based on spatial and spatial-temporal overlap respectively. The results do indeed show a more monotonic relationship than linear relationship, which explains the higher coefficient values for Spearman's rank correlation coefficient. Moreover, the log-log plot for the spatial-temporal overlap shows a higher coefficient of determination $R^2 = 0.74$ than the square of the correlation coefficients, which is an evidence for the non-linear nature of the relationship.

The members of a cohesive groups have most of their ties inside the group, and

	r	ρ	$p(\epsilon)$
Scos:	0.45	0.6	0
WSCos-Dens:	0.43	0.6	0
WSCos-Dist:	0.5	0.63	0
WSCos-User:	0.46	0.67	0
WSCos-Entr:	0.51	0.65	0
Scol:	0.15	0.43	0.0006
Col:	0.22	0.23	0.0562
s:	0.48	0.63	0
s-Dens:	0.57	0.73	0
s-Dist:	0.66	0.81	0
s-User:	0.6	0.78	0
s-Entr:	0.58	0.79	0
s-Extra Role:	0.32	0.75	0

Table 4.9: The correlation between 2-plex cohesion measurement calculated according to equation 4.2 and all mobile homophily measurements. The last column contains the p-value of the corresponding Spearman's correlation coefficient. The p-value indicates the probability that social proximity and mobile homophily have no relationship (page 11).

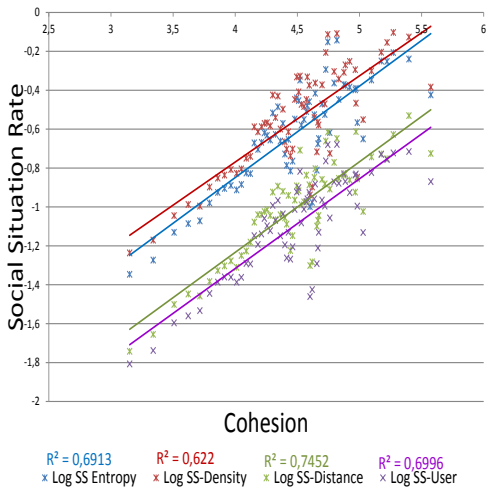

Figure 4.5: The correlation between the cohesion measurement according to equation 4.2 and different weighted social situation rates.
The correlation between the cohesion measurement according to equation 4.2 and different weighted social situation rates.

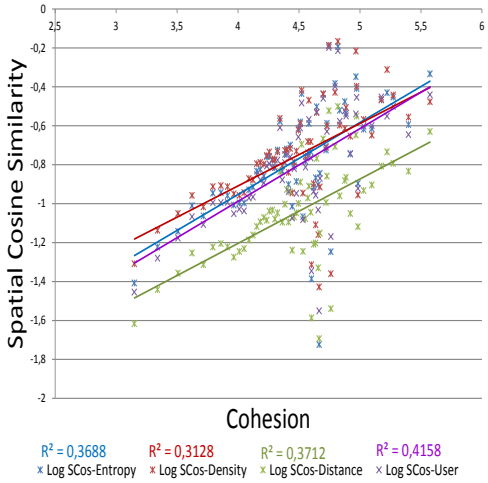

Figure 4.6: A log-log plot representing the distribution of cohesion measurement according to equation 4.2 with respect to different weighted spatial cosine similarities.
A log-log plot representing the distribution of cohesion measurement according to equation 4.2 with respect to different weighted social situation rates.

most cohesive groups are of a small size (≤ 6). This accords perfectly with the strong negative correlation between group cohesion and both common neighbors and Adamic & Adair (table 4.10), because the social proximity measures rather are low and the cohesion very high. The positive correlations between group cohesion on

one hand and both Jaccard's coefficient and the degree of cliquishness on the other hand confirm that most cohesive groups have most of their ties inside the group (table 4.10). The most cohesive groups have most of their ties inside the group, which means higher degree of cliquishness and a higher jaccard coefficient, thus the positive correlation coefficients.

	r	ρ	$p(\epsilon)$
CN:	-0.34	-0.86	0
AA:	-0.42	-0.75	0
Jacc:	0.06	0.2	0.095
DoC:	0.37	0.44	0

Table 4.10: The correlation between 2-plex cohesion measurement calculated according to equation 4.2 and all social network measurements

Finally it is worth mentioning, that the weighting factors used all led to higher correlation coefficients, with "distance from home" proving to be the best weighting factor. Although the results show that people tend to form ties with people in their neighborhood, a meeting between two users living distant from each other is an indicator for the importance of their tie, because as stated earlier, people are rather willing to cover greater distances, mostly in order to meet close/important friends.

Chapter 5

Causation Effect Between Social Network and Mobility Prediction

The location history of a user is not sufficient for explaining the whole mobility behavior of the user. Human beings are members of a social network. They influence and are influenced by other members of the social network. The social relationships of a user can be coarsely divided into two types, namely close friends (strong ties) and acquaintances (weak ties). The user and their close friends exhibit some kind of homogeneity in their beliefs, thoughts, goals, emotional needs, movements etc. resulting in high interaction and spatial-temporal overlap. The user interacts less and spends less times with their acquaintances, hence their goals, thoughts, emotional needs, movements, etc. exhibit considerably higher heterogeneity, thus, acquaintances are important for transmitting more novel information between different social communities. The focus of this chapter is on extending the ST PPM VOMM mobility model for integrating influences from the social network, in order to enhance the prediction accuracy.

Chapter summary: Section I contains a short introduction to the causation effect. Section II presents a few social influence models that can improve prediction of the effects resulting from preceding causes are presented. Section III looks at the integration of both synchronous social influences and general social trends into the ST PPM VOMM. This section also includes an introduction to the drift function which describes the decay of social network influence. Due to extensive reasons the location history of a user may not be complete, section IV contains a method for handling missing data. Section V contains the empirical results based on the Foursquare dataset, in order to show how social influences flow inside and between different social groups. Further, the section contains an in-depth correlation analysis between improvement in accuracy and mobility measurements, in order to emphasize the importance of social networks in explaining the movement behavior of an individual.

5.1 Introduction

The correlation between mobile homophily and social proximity was discussed in the previous chapter, but correlation is not the same as causation. Correlation shows a statistical dependence between two events rather than a relationship of cause and effect. The order of events in correlation analysis is of little importance, but is far more significant in causation effect, because causation requires the cause to precede or coincide with the consequence.

In general, three types of causation effect can be distinguished, namely necessary (a cloud is necessary for rainfall), sufficient (wind is a sufficient cause for the rustling of the trees) and contributory. A cause is contributory **"If the presumed cause precedes the effect, and altering the cause alters the effect. It does not require that all those subjects which possess the contributory cause experience the effect. It does not require that all those subjects which are free of the contributory cause be free of the effect. In other words, a contributory cause may be neither necessary nor sufficient, but it must be contributory"** [Riegelman, 1979]. John Leslie Mackie introduced the INUS condition for contributory causes in [Mackie, 1974], the INUS condition refers to **"insufficient but non-redundant parts of a condition which is itself unnecessary but sufficient for the occurrence of the effect"** [Mackie, 1974].

The influence of the mobility of users on the mobility of their neighbors is a contributory cause. This chapter describes an investigation into whether a visit(cause) of a user(influencer) to a location, can cause (contribute to) a visit(effect) of friends(followers) to the same location, either the cause precedes the effect(asynchronous influence, i.e. the influencer is absent during the visit of the friends) or it coincides with it (synchronous influence, i.e. the influencer accompanies the follower during their visit). The term social influence indicates the causation effect between influencers, from whom the influence originates, and followers. From the perspective of a mobile user u_i, an influencer is another mobile user u_j, whose mobility influences the mobility of the user u_i and the two users are socially connected, i.e. a tie exists between the two users $e_{ij} \in E$ on the social network graph G.

5.2 Social Influence Models

Social or group influence has been studied for several decades and is defined as the behavior of an influencer (entity/user/item) at time t_i which can influence the behavior of a follower (another entity/user/item) at time t_j where $i \leq j$ [Pan et al., 2012]. A model that incorporates the behavior of the influencers in a predictive model and consequently improves prediction of the behavior of the follower is called a social influence model. A few social influence models are introduced below.

5.2.1 Periodic & Social Mobility Model

[Cho et al., 2011] proposed a Periodic Mobility Model (PMM) for predicting the locations of an individual. The Author uses data collected from two online location-

based social network platforms (Gowalla and Brightkite), and the traces of 2,000,000 mobile phone users from a European country for evaluating their mobility model. From their analysis they found that people return to places they've visited earlier with a probability of 53% and visit the home of a friend with a probability of 63%. Based on the intuition that most human movements are centered around work and home, they built a Periodic Mobility Model (PMM) to consist of two states only, namely "home" and "work". PMM models the mobility of a user as a time-variant stochastic process, where the temporal dynamics of human mobility are captured based on a day-specific periodic transition model. To extract the geographical location of the "home" and "work" states, they subdivide the space into cells of 25 km. The geographical position of each of the two states is calculated based on the distribution of check-ins within the 25 km cells. After inferring the centers of the two states, PMM uses a mixed Gaussian distribution centered around the two states for predicting the future location of the user. PMM uses equation 5.1 for predicting the location of the user.

$$\begin{aligned} P(x_t = x) &= p(x_t = x|c_u(t) = H)P(c_u(t) = H) \\ &+ p(x_t = x|c_u(t) = W)P(c_u(t) = W) \end{aligned} \tag{5.1}$$

where H denotes the "home" and W the "work" state of the model and $c_u(t)$ is a function of the time of the day that can calculate the probability of the model being in one of the two states. Based on the Shannon entropy for each hour of the week, they found that the entropy during the evenings of the work days and during the weekends is higher, which means that the mobility of people during these time periods is less predictable. The Periodic & Social Mobility Model (PSMM) is an extension of the PMM model integrating the effect of social influence on mobility. Check-ins during workday evenings and during weekends are denoted as social check-ins. Assuming a user is making a social check-in, the PSMM calculates the social influence and adds it to the result of the PMM. Social influence is calculated as a power law of two factors, namely the time and the location of the check-in of each friend of the user. The researchers' experimental results showed that PMM can predict the location of a user with an accuracy of 42% with an average relative distance error of 2.9%. The social model PSMM can predict the location of a user with a similar probability to PMM, but improves the average relative distance error to 2.7%, which corresponds to an improvement by 10%. Further they took into account the population density in urban and rural areas and found that people tend to meet within a radius of "reach" of less than 100 km from their homes. They found that the influence of the mobility of distant friends is higher than that of those friends who live in the neighborhood. PSMM is a very simple model having only two states, the average relative distance error of 2.7% for a movement over $10km$ is equivalent to an absolute error of 270 meters from the exact location, which is indeed a significant error.

5.2.2 Random Utility Decision Models (RUM)

Random Utility Decision Models (RUM) (or discrete choice models) are statistical procedures that predict the choice of a user among alternatives based on a utility

function [Mcfadden, 2001], which can take many factors into account. [Hackney and Axhausen, 2006] proposed a work investigating the interdependence between social network and travel behavior based on RUM. It models a user's utility for a location j based on both the travel cost from a start location to a destination location j and the social influence of the destination location. The social influence of a location is calculated based on factors such as the number of friends at the location and/or number of friends-of-friends. The model is able to insert and update new ties based on the presence of co-locations. The model is also able to remove ties between members of the social network if the last co-location of two users is older than a threshold. The authors do not consider any individual or geographical characteristics, their focus is only on social travel. The methods are evaluated based on simulated data, therefore it is not clear how the model will behave when using real life data

5.2.3 Topical Affinity Propagation (TAP)

[Tang et al., 2009] has analyzed topical influence and proposed an influence model called Topical Affinity Propagation (TAP). For each topic they determined a set of representative members of a social network and investigated the influence of a member on their friends. The proposed Topical Factor Graph (TFG) incorporates both user-specific topic distribution and network structure into one probabilistic model. For a social network $G = (V, E)$ and a topic s they define a Topical Factor Graph $G_s = (V_s, E_s)$ for all topics $1 \geq s \geq T$. They use the Topical Affinity Propagation (TAP) learning for learning the model parameters. This method is based on the sum-product algorithm described in [Kschischang et al., 2001]. Their experimental results show the success of TAP in identifying topic-based influence between members of a social network in large-scale databases. The topics are assumed to be independent, ignoring any kind of dependencies, which usually exist between locations if the model is applied to the mobility of human beings. In addition, the interest of a person in a location is highly dependent on the time (daily routines), certain places are visited only during certain times. Temporal dynamics and the temporal dependencies are not considered in this work either. Finally, the tie between connected members of the network is integrated using a binary variable that ignores the tie strength between friends.

5.2.4 Influence Models Based on DBN

A group mobility model has been proposed in [Musolesi et al., 2004] which takes into account the social ties between members of the group (in contrast to random group mobility models [Camp et al., 2002]). Based on the DBN approach, the authors investigated the influence of groups on individual human movement. They distinguish between two types of groups, namely the social and the geographical group. A social group represents a set of connected members in the social network, whereas a geographical group is formed and dispersed more dynamically at a certain time and location (such as home or work groups). The users in a social group are socially connected and potentially can be co-located, whereas geographical groups contain a set of users, who are co-located at a certain time and location, but are not

necessarily socially connected. An individual moves within the sphere of influence of a geographical group at any point in time, until they decide either to move between the groups or leave all the groups and move independently. An individual joins a group based on the attractiveness of the group and the difficulty of reaching that group. The attractiveness of a group for a user u_i is modeled using an interaction matrix according to the past co-locations shared by this user and each member of the group. Each element of the matrix e_{ij} is an interaction factor e_{ij} between the two users u_i and u_j. Each user u_i has a sociability factor SF_{u_i}, which indicates their attitude towards a group. The authors used a social threshold of 0.25 while calculating SF_{u_i} and considered only those members of the group who had an interaction factor e_{ij} higher than this threshold. The difficulty of reaching a group is considered in the model as a function of physical distance. The experimental results show a strong influence of groups on individuals moving in and between these groups. Unfortunately the authors evaluated their model based on (possibly unrealistic) artificially generated social relationships among mobile phone users.

[Sadilek et al., 2012] used the twitter API to collect publicly available tweets from two distinct areas (LA and NY). They collected all the tweets within a time period of one month. They considered only the users who had at least 100 GPS-tagged tweets during this one month. The authors introduced a lot of restrictions by their choice of locations and friendships. For example, they considered only those locations that were visited at least five times by a user. Further, two users were only considered friends if they "did reciprocate following" (reciprocate following means two users follow each other on Twitter), i.e. only strong ties were considered. Furthermore, they considered only n friends of each user and selected n from the range $0 - 9$. Taking into account both temporal and social dependencies in a location prediction model based on a DBN approach, they could predict the future location of a user with an accuracy of approximately 84%. The reason for the high performance lies in the restrictions they made. The locations are coarse as all the locations within 100 meter are clustered to a single unique location. For each user, locations having less than 5 visits are ignored, which means that only frequently visited locations were considered. Another restriction is the short 20 minute time interval they use, which is not really far into the future. In addition, they use an advantageous accuracy calculation. Let us consider a user who has visited four locations. Let us suppose the user spent ten hours at home, eight hours at work and the rest of the 24 hours at two other locations. Calculating the performance of a predictor by considering only real movements between locations as true predictions, the predictor reaches an accuracy of $\geq$ %50, whereas predicting every 20 minutes and counting it as true prediction, even if the user the user does not move, the predictor reaches an accuracy of $\geq 75\%$. On the network side, they only consider users who "do reciprocate following", which means that mutual influence is considerably higher.

[Pan et al., 2012] have proposed a general social influence model that can be applied to any interaction network in any social system (including social networks like Foursquare, Facebook, etc.). Their general social influence model is based on a simple mixture approach with fewer parameters than the Hidden Markov models called the dynamical influence model. The model tries to find the conditional property $P(x_t^{(u_i)}|x_{t-1}^{(u_1)}, ..., x_{t-1}^{(u_j)})$ indicating the influence of the previous states of neighbors

on the current state of an individual u_i as follows:

$$P(x_t^{(u_i)}|x_{t-1}^{(u_1)},....,x_{t-1}^{(u_{|N(i)|})}) = \sum_{u_j \in N(i)} \mathfrak{T}_{u_i,u_j} \times \mathbf{I}(x_t^{(u_i)}|x_{t-1}^{(u_j)}) \tag{5.2}$$

where $N(i)$ is the set containing the neighbors of the user u_i, $\mathfrak{T}$ is the (weighted) adjacency matrix capturing the tie strength between the users in the interaction network, $\mathbf{I}$ is a matrix that captures the influence from the previous states of the neighbors $x_{t-1}^{(u_j)}, u_j \in N(i)$ to the user u_i, over the current state of that user $x_t^{(u_i)}$.

In the above social influence model, the tie strength matrix remains the same over time. Due to the existence of extensive evidence that tie strength may change over time, the authors extended the static model. Instead of using only one static tie strength matrix $\mathfrak{T}$, the extended model uses a set of different tie strength matrices $\{\mathfrak{T}^{(1)},...,\mathfrak{T}^{(k)}\}$ for capturing dynamic changes over time (in a brain storming scenario, two users may interact differently over time), where k is a hyper-parameter set by the authors for controlling the number of tie strength matrices. The authors use a switching latent state variable $r_t \in \{1,2,...,K\}$ which controls the current tie strength matrix to be used among a set $\mathfrak{T}$ of K tie strength matrices, equation 5.2 is then changed to:

$$P(x_t^{(u_i)}|x_{t-1}^{(u_1)},....,x_{t-1}^{(u_{|N(i)|})}) = \sum_{u_j \in N(i)} \mathfrak{T}(r_t)_{u_i,u_j} \times \mathbf{I}(x_t^{(u_i)}|x_{t-1}^{(u_j)}) \tag{5.3}$$

where $r_t \in \{1,2,...,K\}$ indicates the current tie strength matrix $\mathfrak{T}(r_t)$ and t is the current time step, for more details please refer to [Pan et al., 2012]. The model is evaluated on three different scenarios; the most interesting scenario is the prediction of turn-taking in discussions. Each user in the model used a wearable sociometric badge, "a wearable electronic device capable of automatically measuring the amount of face-to-face interaction, conversational time, physical proximity to other people, and physical activity levels using social signals derived from vocal features, body motion, and relative location" [soc, 2013]. The model tries to predict when a user starts and ceases speaking using only the audio volume variance. Each user has two hidden states, "speaking" and "not speaking". The dynamical influence model should capture how the states of a user are influenced by the states of their neighbors. The model considers only synchronous influence, because all the neighbors are present at the same location at any time step. The evaluation results are promising and show an accuracy improvement of up to 25% compared to human guesswork and up to 40% compared to random guessing. However, it is questionable how the dynamic influence model will perform when applied to a huge social network with some users having thousands of friends and a state space that covers ten thousands of locations plus temporal features, additional contexts and discrete knowledge from external information sources (such as calendar entries).

5.3 Integration of Social Context

The ST VOMM PPM mobility model introduced in the previous chapters incorporates features related to both the spatial and temporal context of a mobile user. The movement of human beings is subject to social influences. The mobility of one user can be influenced by the mobility of other users from the same community (such as their circle of friends). The Socio-Spatial-Temporal (SOST) PPM VOMM is an improved ST PMM VOMM that incorporates social influence factors in order to improve the accuracy of next location prediction. The influence factors arising from the social network on the mobility behavior of an individual are of different types, generally, we distinguish between two types of social influence, namely synchronous specific and general social (trends) influences.

A Synchronous specific influence factor represents the cases when a user and a set of their friends are involved in the same social situation. We refer to the friends from whom the influence originates as influencers. An example of synchronous specific social influence factor is a social situation, in which one or more friends visit a restaurant they have visited earlier, either together or with other friends (because they like the restaurant). The current visit to the restaurant is a social situation, which is synchronously influenced by past visits of the friends to the same restaurant, either in the form of social situations or single visits. We refer to past visits of friends as social influence factors. Thus, synchronous social influence has two preconditions, first the user must be currently involved in a social situation, second the availability of location histories of friends. The introduce Synchronous specific influence factors in more details in the next subsection.

General social influence factors represent the general movement patterns and social trends in a user's community, for example the favorite bar or club of their circle of friends, a hip new restaurant in the city or an inexpensive shopping mall, etc. The users in the same circle of friend or the same community share common interests, hobbies, thoughts, beliefs, etc. which precipitate interest in common locations, or similar movement behavior in the same spatial-temporal context. For general social trends, the users are not necessarily together at the same time. The influencers of general social influence factors are not (synchronously) observable, the influence is more likely to be transmitted in the form of recommendations through media like the phone, e-mail, SMS, online social networking platforms, etc. or may simply be a general mobility pattern in the community. General social influence has only one precondition, namely the availability of location histories of friends.

5.3.1 Synchronous Specific Social Influence

Synchronous specific social influence occurs when a user is currently involved in a social situation. We refer to the friends in the social situation as influencers. As stated earlier, a precondition for synchronous specific social influence to take place is the involvement of the user in a social situation, thus the detection of social situations is very important. We assume that the current social situation of the user is known at any point in time [Groh et al., 2010]. The second precondition is the existence of location histories of friends. A past visit of a user to a location is

conditionally not independent from other users being present at the same location within a specific period of time δt from the time of the visit. Therefore we group the visits of friends U to the same location within δt to a social situation.

The value of δt must be well considered, because on one hand choosing a high value for δt increases the uncertainty of detecting the correct social context of a visit, on the other hand choosing a low value for δt increases the probability of overlooking a valid social situation. The correlation analysis between mobile homophily and social proximity from the previous chapter has shown a moderate to a strong correlation when setting δt to one hour, therefore we keep on setting δt to one hour.

The Social context of a visit to a location from the point of view of a user u_i is a tuple $\mathfrak{s} =< U, q, \lambda, t, c >$, where U is the set of users present at location q representing a synchronous specific social influence factor, λ represents the set of temporal features (such as work day or weekend, day of week, hours of day etc.) extracted from the time stamp t of the visit and c is a counter that bookkeeps the occurrence of that specific social influence factor. Let user u_i be the user whose next location is going to be predicted and $N(i)$ the neighbors (friends) of u_i, then U is a subset of $U \subseteq N(i) \cup u_i$. According to the users present in U, the social context of a visit can be categorized into three different classes of synchronous specific social influence factors $\mathfrak{s}$:

- **Class I social influence factors -** is a social situation that contains the user u_i and at least one of their neighbors, i.e. u_i is visiting a location with at least one of their friends.
- **Class II social influence factors -** is a social situation that contains at least two neighbors of the user u_i without the presence of u_i in the social situation.
- **Class III individuals based social influence factors -** Contains single visits of the neighbors without the presence of other users.

The inclusion of class II & III social influence factors injects a vast amount of extra knowledge into the mobility model of an individual. It helps predict locations which have been visited by friends even if the user themselves has never been there before. The prediction of locations where the user has never been before is almost impossible if only the location history of the individual user is available (but may be possible to a limited extent if additional data sources such as their personal calendar are considered).

Social influences arising from past visits of friends to a location q depend on the current social situation of the user and the temporal features λ extracted from the current time step t. The probability a user visits a location under consideration of the spatial-temporal context and synchronous specific social influence factors is captured by the conditional probability $P(q|U, s, \lambda)$, where s is the current spatial context, λ the set of current temporal features and U contains the users involved in the current social situation. We assume the current social situation to be independent from the spatial context except from the current location, i.e. previously visited locations have no influence on the current social situation. This assumption seems reliable, because people usually meet at a location and only in rare cases visit multiple locations

sequentially. Even though, sequential visits to multiple locations with friends may be a trend for a particular circle of friends. Such generalities and trends are subjects of the next section about general social influences. We assume that the probability mass $P(q|U, s, \lambda)$ can be estimated according to equation (5.4):

$$P(q|U, s, \lambda) = \underbrace{P(q|U, \lambda)}_{\text{Social influence}} * \underbrace{P(q|s, \lambda)}_{\text{Individual mobility}} \tag{5.4}$$

Where λ contains the set of temporal features, U the current social situation and s the current spatial context of the user u_i in question for location prediction. Estimation of the probability mass $P(q|U, s, \lambda)$ is the multiplication of two terms. The right term represents the probability of visiting a location given both spatial and temporal contexts, which can be estimated from the individual location history as in the previous chapters. The left term represents influences arising from the current social situation and temporal context of the user.

5.3.1.1 Social Situation Detection

Detection of a social situation is a key operation of the SOST PPM VOMM model. SOST PPM VOMM detects social situations based on spatial-temporal overlaps among the movements of two neighbors (friends). Thus, the detection of social situations is closely related to the underlying sensor data and requires comparison of the location data of different users. Specifically the detection of a class II & III social influence factors depends on the availability of global location identifiers. The Foursquare dataset contains global identifiers for the locations making social situations easy to detect, whereas the Reality Mining dataset [Eagle and Pentland, 2006] contains GSM cell tower IDs of the mobile phone provider. The users in the Reality Mining dataset are responsible for labeling the cell tower IDs at the significant locations they have visited. Unfortunately the labels are very user-specific so comparison of labels used by different users is impossible, because each user labels the same cell tower ID differently. The Foursquare dataset on the other hand has the disadvantage of not containing sensor data (such as device-to-device communication) that can indicate whether two users have interacted. Therefore, SOST PPM VOMM assumes two users were involved in the same social situation if they check-in at the same location within a time period of one hour. SOST PPM VOMM proceeds to detect social influence factors of the three different classes as follows:

- **Class I social influence factors:** For each Foursquare check-in $< u, q, t >$ of a user, SOST PPM VOMM searches their friends' histories for check-ins at the same location within a time period of one hour. The Reality Mining dataset contains user-unique IDs for the mobile phone cell towers, but since the cells are relatively large (3x3 KM), and since the mobile device can be observed by multiple cell towers even if the user u_i stays at exactly the same location, SOST PPM VOMM instead uses device-to-device communication via Bluetooth to detect social influence factors. Each time two friends communicate over their devices SOST PPM VOMM considers it a social situation and looks at each user's individual data for their current location.

- **Class II social influence factors:** SOST PPM VOMM iterates through the friends of the user and proceeds similarly for detection of class I social influence factors looking for class II social influence factors. The only difference is the absence of the user u_i in the social situation. As stated earlier, class II social influence factors can evolve at locations where the user u_i has never been before, therefore their detection helps to add a huge amount of valuable information that cannot be detected from the individual's location history of u_i.
- **Class III individual-based social influence factors:** The remaining (unassigned) check-ins of the friends are considered social influence factors with only one user.

Unfortunately it is not possible to detect class II & III social influence factors in the Reality Mining dataset, because the user-specific labels do not allow a global comparison of the locations of different users.

5.3.1.2 SOST PPM VOMM Tree

The standard PPM VOMM uses the number symbols $\Sigma_s|$ appearing after a context s in the equations (2.11) and (2.12). Inclusion of class II and class III social influence factors causes the number of symbols in the alphabet of the model and the number of symbols appearing after a context to increase rapidly, therefore it affects the whole probability distribution and consequently has a significant effect on the performance of the individual mobility model. In order not to decrease the accuracy of the individual mobility models of the users, we add a new tree for managing the influence the social factors have on the mobility of the user. We refer to the new Socio-Spatial-Temporal tree as the SOST PPM VOMM tree. We use the SOST PPM VOMM tree to calculate the conditional probability $P(q|U, \lambda)$ in equation (5.4).

Table (5.1) shows the features and their domains, that are used in the SOST PPM VOMM tree.

Variables	Domain	Description
Σ_{loc}	$\{l_1, l_2, ..., l_i\}$	The set of locations visited by the user
W	$\{Wd, We\}$	a binary variable representing whether it is a weekend day or a work day
D	$\{Sun, Mon, ..., Sat\}$	The day of week
$S^{\Delta t}$	$\{S_1, S_2, ..., S_j\}$	The number of time slots calculated by dividing the hours of day by Δt, setting $\Delta t = 1$ means that each hour of day represents a slot
U	$\{u_1, u_2, ..., u_j\}$	The set of users present in the current social situation

Table 5.1: The features included in the SOST PPM VOMM model.

The nodes of both PPM VOMM tree shown on figure (2.7) and ST PPM VOMM tree shown on figure (3.1) corresponds to locations visited by a user u_i and temporal

features related to the times user u_i visited these locations respectively. Each node of both trees manages a counter for bookkeeping the number of its occurrences in the location history of the user u_i. The nodes of the SOST PPM VOMM tree immediately under the root node are labeled with locations and nodes at deeper levels are labeled with temporal features contained in the location histories of the user u_i and their friends $N(i)$. The nodes of SOST PPM VOMM tree manage a set of tuples of form $< U, t, c >$ instead of single counters. Each tuple corresponds to a social influence factor evolved at a location q and temporal context λ represented by a node of the SOST PMM VOMM tree. Each tuple consists of a set of users U, the time stamp t of the last occurrence and a counter for bookkeeping the number of occurrences of a social influence factor up to time step t. Each path in the SOST PPM VOMM tree corresponds to a location q and a temporal context λ contained in the location histories of a user u_i and their friends $N(i)$. Figure 5.1 shows an example SOST PMM VOMM tree and zooms into a node in order to illustrate how SOST PMM VOMM manages different social influence factors.

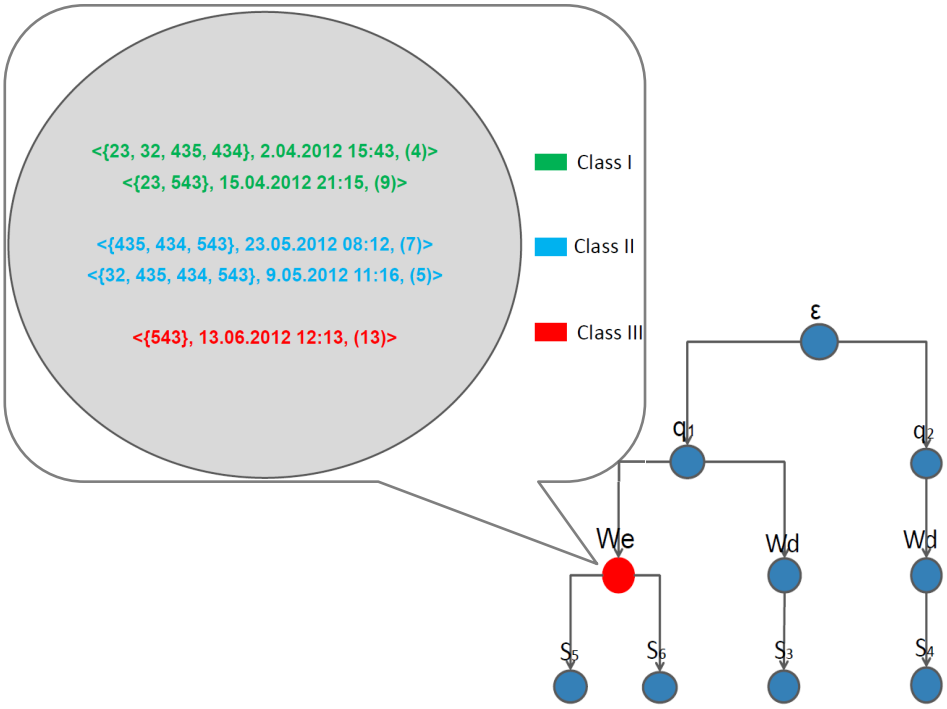

Figure 5.1: An example SOST PPM VOMM tree of a user with ID = 23, the nodes immediately under the root node are labeled with locations, the nodes at deeper levels are labeled with temporal features such as work and weekend days and time slots of day. Unlike the PPM VOMM tree on figure (2.7), each node in the SOST PPM VOMM tree has multiple tuples for managing the occurrence of social influence factors. The figure zooms into the red node in order to illustrate how SOST PPM VOMM manages the three classes of social influence factors at location q_1 on weekend days. Each social influence factor is a tuple consisting of a set of users who build together the social influence factor (the numbers in curly braces represent the IDs of the users), the time stamp of its latest occurrence and a counter (the number in parenthesis) for bookkeeping the number of its occurrences. Social influence factors of different classes are colored differently.

Once a social influence factor is identified, SOST PPM VOMM creates a key by

simply sorting the user IDs and concatenating them using a separator (semicolon). For simplicity reasons, we use the set of users U involved in a social influence factor to refer to a synchronous specific social influence factor. We integrate social influence factors into SOST PPM VOMM tree as follows:

We traverse the SOST PPM VOMM tree to find a corresponding node according to the current location q and temporal context λ of each of the social influence factors. We insert a new path if a corresponding node does not yet exist in the tree (necessary for class II & III social influence factors). We initialize a new counter for a social influence factor that occurs for the first time. The integration of social context into our mobility model simply corresponds to adding new paths to the SOST VOMM PMM tree, incrementing or initializing counters for each different social influence factor.

The elegance of this method lies in the simplicity of incorporation of a huge amount of additional social knowledge into the individual mobility model of the specific user u_i. Hence the inclusion of class II and class III social influence factors merely requires adjustment of the prediction tree instead of a multiplicatively growing transition matrix as observed in the naive Markov model implementations [Bapierre et al., 2011].

5.3.1.3 Synchronous Specific Social Influence Estimation

SOST PPM VOMM has to estimate the social influence part $P(q|U, \lambda)$ of equation (5.4) in order to predict the future location of a user using synchronous specific social influence factors and both spatial and temporal features. Unlike the temporal features, social influence factors do not support any inclusion semantic. For example, a social influence factor is not necessarily included in another social influence factor. Further, social influence factors do not have an order of occurrence such as consequentially visited locations, i.e. different social influence factors do not occur in a specific order. Therefore, it is not possible to use the standard escape mechanism of PPM VOMM and use a more general social influence factor when the current social influence factor is not yet seen in the location histories of the users. SOST PPM VOMM makes use of the overlap between two social influence factors as a similarity measure in order to compare two social influence factors. The Jaccard Similarity Coefficient (JSC) [Jac, 2013] is a measure for calculating the similarity or diversity between two sample sets. JSC is defined as the size of the intersection divided by the size of the union of the two sample sets (5.5):

$$Jacc(U, \acute{U}) = \frac{U \cap \acute{U}}{U \cup \acute{U}} \tag{5.5}$$

SOST PPM VOMM uses JSC for comparing two social influence factors. The use of social situation (social influence factor with multiple users) has the advantage of capturing the conditional dependence between visits of different users involved in the same social situations. The combination of social influence factors and JSC helps determine the amount of influence arising from past synchronous specific social influence factors at a location on the current visit of the user to that location using the current social situation of the user as follows:

Let U be the set of users involved in the current social situation, further let η be a node that corresponds to a spatial-temporal context (a location q and the set of temporal features λ extracted for the prediction time). Equation 5.6 calculates the impact of synchronous specific social influence factors by modifying the counter of the node η:

$$C_{\mathfrak{s}=(U,\lambda,q)} = \sum_{\acute{U}} C_{\mathfrak{s}=(\acute{U},\lambda,q)} * Jacc(U,\acute{U}) \tag{5.6}$$

where $C_{\mathfrak{s}=(\acute{U},\lambda,q)}$ is a counter that hold the number of occurrences of a synchronous specific social influence factor with a set of users $\acute{U}$ at location q and temporal context λ. $C_{\mathfrak{s}=(U,\lambda,q)}$ is an estimate of the occurrence of the current social situation with a set of users U according to the spatial-temporal context (q,λ). SOST PPM VOMM makes use of equation (5.6) for estimating the probability mass $P(q|U,\lambda)$. Let ς be the set containing a node $\eta = (q,\lambda)$ corresponding to a location q and temporal context λ and its ancestors, equation (5.7) estimates the probability mass $P(q|U,\lambda)$:

$$P(q|U,\lambda) = \frac{C_{\mathfrak{s}=(U,\eta)}}{|\Sigma_\varsigma| + \Sigma_{\eta' \in \Sigma_\varsigma} C_{\mathfrak{s}=(U,\eta')}} \tag{5.7}$$

Equation (5.8) represents another possibility to estimate the probability mass $P(q|U,\lambda)$ by changing the denominator of equation (5.8)

$$P(q|U,\lambda) = \frac{C_{\mathfrak{s}=(U,\eta)}}{\sum_{\acute{U}} C_{\mathfrak{s}=(\acute{U},\eta)}} \tag{5.8}$$

The denominator in equation (5.8) is smaller than the denominator in equation (5.7), because it depends only on the node η and not its ancestors. Thus equation (5.8) gives synchronous specific social influence factors more importance than equation (5.7).

The mobility model behaves like an individual mobility model when the current social context contains only the current user u_i, because the individual movements of the user will have the highest possible Jaccard coefficient of one and in most cases a higher number of occurrence compared to other locations with social situations containing the user u_i. All social influence factors that do not contain the user will have a Jaccard coefficient of zero, thus will have no effect on the prediction of the future location of the user u_i.

5.3.1.4 The Impact of Tie Strength

Equation 5.5 weights the social ties of the user u_i to their friends who are involved in synchronous specific social influence factors equally. Each user ranks their social ties according to the importance of their relationship to each of their neighbor. The strength of a tie is not easy to classify, because according to [Krackhardt, 1992], the definition of tie strength is tightly related to subjective criteria like emotional

intensity and intimacy. Nevertheless, a strong tie is a relationship that meets three necessary and sufficient conditions, namely **interaction:** both users must interact with each other, **time:** both users must have a history of interaction over a long period of time, and finally **affection:** both users must feel affection for one another [Krackhardt, 1992]. Although affection feeling is hard to be quantified, but interaction can be measured using the amount of communication, the amount of shared information, the amount of spatial-temporal overlap between their movements, the duration of time spent together, etc. Due to the lack of more information, we simply assume (even if insufficient), that, the strength of a tie can be quantified using the amount of spatial or spatial-temporal overlap between their movements, following the intuition that close friends have more locations in common compared to distant friends. In order to increase the adaptability of the model, we modify equation (5.5) as follows:

$$\text{Jacc}(U, \acute{U}) = \frac{\sum_{u_j \in U \cap \acute{U}} \mathfrak{T}(u_i, u_j)}{\sum_{u_k \in U \cup \acute{U}} \mathfrak{T}(u_i, u_k)} \tag{5.9}$$

where $\mathfrak{T}(u_i, u_j)$ is the tie strength between the users u_i and u_j.

The tie strength between two users can be calculated using either spatial, or spatial-temporal overlap between the mobility of two users (see section 4.4). The correlation analysis between network proximity and mobile homophily has shown a higher correlation coefficient when calculating mobile homophily using measurements based on spatial-temporal overlap. Nevertheless, using measurements based on spatial-temporal overlap unfortunately has the disadvantage of ignoring the influence of visits of friends prior to the first social situation involving both users. In other words spatial-temporal overlap means calculating tie strength considering only class I social influence factors, whereas spatial overlap means using all three classes of social influence factors.

Let $N(i)$ be the set of neighbors of user u_i, SOST VOMM PPM calculates tie strength based on spatial overlap only between user u_i and a friend from the set $N(i)$ according to equation 5.10:

$$\mathfrak{T}(u_i, u_j) = \frac{\sum_{l \in L} col_l(u_j)}{\sum_{u_k \in N(i)} \sum_{l \in L} col_l(u_k)} \tag{5.10}$$

where L is the set of locations visited by the user u_i, $col_l(u_j)$ is the number of visits of user u_j to location $l \in L$. Note that the tie strength according to equation 5.10 is not commutative, i.e. $\mathfrak{T}(u_i, u_j) \neq \mathfrak{T}(u_j, u_i)$, because the users have different sets of neighbors, thus the denominator in equation 5.10 differs according to the point of view of each user. The non-commutative tie strength is closer to reality, because (psychologically) each of the two users may value the importance of their relationship differently.

Equation 5.10 does not consider the importance of a location for the users. Different weighting factors 4.4.3 such as the distance from home, the density and the entropy of a location can help calculate a weighted version of tie strength according to equation 5.11.

$$\mathfrak{T}(u_i, u_j) = \frac{\sum_{l \in L} col_l(u_j)\omega(l)}{\sum_{u_k \in N(i)} \sum_{l \in L} col_l(u_k)\omega(l)} \tag{5.11}$$

$\omega(l)$ is a function that weights visits to location l according to one of the weighting factors.

5.3.1.5 Drift Function

People are members of various social groups and communities such as a family, a circle of friends, a working environment. Members of the same group to some extent exhibit similarity in their interests, beliefs, norms, goals, activities, need for emotional closeness, feelings of security, etc. Group affiliation, common goals, norms, activities, interests, etc. produce similarities in the behavior of the members, which are subject to continuous change over time. The changes lead to the formation of new and/or weakening of existing social ties [Thomas and of Sociology, 2009] which in turn leads to changes in the mobility behavior of the users. Therefore, the amount of influence on the mobility of an individual resulting from a synchronous specific social influence factor is highly dynamic and subject to decay over time. Suppose l to be the location of the favorite bar of a circle of friends where the friends meet every Saturday evening. The friends may lose interest in the bar after a while, for example because a new hip bar attracts them more. The drift function weights the influence from a synchronous specific social influence factor on the mobility of a user according to its frequency and the times of its latest occurrence.

We define two different drift functions ψ_1 and ψ_2, namely a power drift function according to equation (5.12) and an exponential drift function according to equation (5.13):

$$\psi_1(U, t) = (1 - \beta)^{(t-t_i)/\Delta t} \tag{5.12}$$

$$\psi_2(U, t) = e^{-\beta(t-t_i)/\Delta t} \tag{5.13}$$

where β is a hyper-parameter that controls the degree of drift and U a synchronous specific social influence factor and $t_i < t$ is the time stamp of the last occurrence of U. The unit of the factor $t - t_i$ is the average stay time Δt in hours. If a user spends on average three hours at a location, the social influence factor decays every three hours by a factor of $1-\beta$ or $e^{-\beta}$ according to both equations (5.12) and (5.13) respectively.

In addition, the drift function ψ leads to a reduction in the impact of rigorous statistics for synchronous specific social influence factors by accelerating their obsolescence. The influence of a social influence factor decays as time goes on and it reaches zero after a certain period of non-occurrence. Note that the hyper-parameter β for synchronous specific social situation factors is different from the hyper-parameter α used for drifting the individual user movements, because social influence factors are highly dynamic and become obsolete faster than individual

movements [Wang et al., 2011, Cho et al., 2011] (individual movements are valid for a longer period of time).

SOST PPM VOMM increments the occurrence of a synchronous specific social influence factor according to equations (5.14):

$$C_U^t(q,\lambda) = C_U^{t_i}(q,\lambda)(\psi(U,t)+1) \tag{5.14}$$

5.3.2 General Social (Trends) Influence

General social influences represent either general trends in the community of the user or influences transmitted asynchronously using other media such as e-mail/phone, for example when a friend recommends a restaurant or a shopping mall to another friend via phone/e-mail, or using online social networking platforms in order to share visits to interesting locations with friends. Contrary to synchronous specific social influence factors, the influencers are not directly observable. The influence arises from an unobservable subset of the community (the neighbors $N(i)$) of the user. Incorporating asynchronous influence helps detect global patterns in the mobility of the neighbors, for example, whenever a user discovers a good restaurant or an inexpensive shop, they recommend it to their friends, who in turn visit the restaurant or the shop with a time delay, which can be a few hours, days or even weeks. Another example would be a pattern such as a visit to location x is always followed by a visit to location y, no matter which user made the visits, for example, in large amusement parks, tourist complexes or when sight-seeing people tend to visit several locations in a specific order according to the geographical conditions of the terrain. A hip club or bar in the community of a user is a further example of general social influence.

Unlike synchronous specific social influence factors, general social influences are difficult to determine because of the lack of reliable evidence to check whether a visit is paid under the influence of previous visits of friends. Furthermore, it is difficult to estimate for how long a visit of a user can influence the future movements of their friends. Therefore, we consider general social influences only in cases where the predicted location has a smaller probability than the probability of an unknown location (the probability $P(escape|s)$ in equation (2.12) after escaping to an empty context $s = \varepsilon$). The extended mobility model integrates the trajectories of all friends in an additional (general) VOMM tree (M') that is constructed by adding the trajectories of all friends as if they were generated by the same user. The extended mobility model predicts the future location of a user u_i in two steps. It uses first the SOST PPM VOMM tree to predict the next location of the user u_i. If the probability of the predicted location is greater than the escape probability, it returns this location and terminates, otherwise it switches to the new general VOMM tree M' in order to predict the next location of the user (equation 5.15)

$$x = \begin{cases} \arg\max\ (P(q|U,s,\lambda)), & P(x|U,s,\lambda) > P(escape|s) \\ \arg\max\ P'(q|s,\lambda)_{M'}, & \text{else} \end{cases} \tag{5.15}$$

5.4 Handling Missing Data

The time elapsing between two consecutive check-ins in online location based social network services (such as Foursquare) exhibits a very high heterogeneity ranging from a few minutes to several days, because check-ins in these services are voluntarily. The heterogeneity can have many reasons such as the users tend to make check-ins at new and interesting locations, which they want to share with their friends rather than reporting every movement they make; a user may forget to make check-ins at some locations or due to privacy issues a user may not want to make a check-in at a private location such as home, work or the home of certain friends [Noulas et al., 2011a]. Therefore it is reasonable to assume that two consecutive check-ins with a temporal lapse exceeding a threshold are not related, i.e. not consequential. In order to decide whether two consecutive check-ins are consequential, SOST PPM VOMM defines a further hyper parameter for the average stay time κ. Two check-ins are assumed to be consequential if the time elapsing between the two check-ins does not exceed κ (this is also applicable when gps traces are available and some measurements are missing). Both training and testing phases are changed in consideration of κ as follows:

At any time step t, the order n of the model is considered as an upper bound of the context length $|s_t|$. During both the training and test phase, SOST PPM VOMM uses a context of a variable length $|s_t| \leq n$. At each time step t, SOST PPM VOMM reads a new symbol of the training sequence and calculates the time elapsed since the occurrence of the previous symbol. Missing data is assumed when the elapsed time exceeds κ. Instead of removing the first symbol of the previous context s_{t-1}, SOST PPM VOMM discards the whole context s_{t-1} and builds a new context s_t that consists only of the new symbol. If the elapsed time is smaller than κ, SOST PPM VOMM builds a new context s_t by appending the new symbol at the end of s_{t-1} and removing the first symbol only if the length of $|s_{t-1}|$ is greater than n. This process is repeated during both the training and test phase until the end of both sequences. The path which should be inserted into the SOST PPM VOMM tree at each time step thus has a variable length with an upper bound of n. The constructed model is now twice variable. It is variable first in the sense of partial matching and it is variable as it switches to a lower order whenever it detects missing data. Using a lower order has the welcome side-effect of having fewer escape cases (a shorter context is escaped in less steps to an empty context), which in turn leads to a reduction in the time needed for processing.

5.5 Empirical Results

We use two datasets for evaluating the performance of SOST PPM VOMM, namely both Reality Mining and Foursquare datasets. Unfortunately the Reality Mining dataset only allows detection of class I social influence factors. It contains 862 social situations, which improves the prediction accuracy for the users who are involved in these social situations by $\approx 1\%$ compared to the pure spatial temporal ST PPM VOMM model. The Reality Mining dataset does not allow an in-depth investigation of the influence of social networks on the mobility behavior of users. In contrary, the

Foursquare dataset allows detection of considerably more social influence factors, which allows the empirical investigation of the influence of social networks on the mobility behavior of humans. The following results are all based on the Foursquare dataset.

5.5.1 Limits of Predictability

The statistics of the Foursquare dataset indicate a very low predictability of the users. Below we define three measures of predictability:

- **Lower bound of Predictability:** An active user can randomly be found on average at one of $|L| = 62.35$ locations, the average entropy of the users is found to be $\mathfrak{E} = 3.48$. A user who visits locations with the same probability has the highest entropy, thus we can determine the average minimum number of locations $|L|$ required for producing an average entropy value of $\mathfrak{E} = 3.48$. The probability of visiting a location will then be $P(l_i) = \frac{1}{|L|}$. The calculation of Shannon entropy is simplified by substituting $P(l_i)$ in equation (4.12) with $\frac{1}{|L|}$ to:

$$\mathfrak{E} = -\sum_{L} \frac{1}{|L|} \ln \frac{1}{|L|} = \ln |L| \tag{5.16}$$

  knowing that value of Shannon entropy $\mathfrak{E} = 3.48$, we can determine the average minimum number of location by $e^{3.48} = 32.46$. A user in the Foursquare dataset is found on average in one of 32.46 different locations, which corresponds to a minimum predictability of $1/32.46 = 3\%$ [Song et al., 2010b]).

- **Upper bound of Predictability:** In almost 38% of cases the users move on to new locations where they have never been before, we assume that these locations are not predictable. Further the users make on average 2.04 check-ins at a location. This means the users visit the remaining 62% of locations with a frequency of 2.68 ($0.38 + x * 0.62 = 2.04, x = 2.68$). Any mobility model will need at least one of the 2.68 check-ins for training. Assuming that a model can predict the remaining locations, the maximum accuracy of $1.68/2.68 = 0.63$ can be achieved for 62% of the check-ins. The mobility of users in the Foursquare dataset is predictable to an upper bound of $0.63 * 0.62 = 39\%$.

- Unfortunately the dataset does not provide us with the amount of time spent at any of the locations, thus we are not able to calculate the predictability of the user's mobility in more detail (in accordance with the limits of predictability by [Song et al., 2010b]). Nonetheless, the focus of next location prediction is on real movement to a new location, thus the contribution of a correctly predicted location is the same for all locations independently of stay time. Thus we can still make use of Fano's inequality ([Fano, 1961] as cited by [Song et al., 2010b]) to calculate the predictability Π^{max} of the mobility of the users:

$$\mathfrak{E} = H(\Pi^{max}) + (1 - \Pi^{max}) \ln(|L| - 1) \tag{5.17}$$

$$H(\Pi^{max}) = -\Pi^{max} * \ln(\Pi^{max}) - (1 - \Pi^{max}) * \ln(1 - \Pi^{max}) \tag{5.18}$$

where $\mathfrak{E}$ is the entropy, $|L|$ is the number of locations and Π^{max} is the predictability of the mobility of a user. Note: in we use natural logarithm in equation (5.18) because we calculated Shannon entropy using natural logarithm in equation (4.12). The average value of Π^{max} over all users is found to be less than 29%. Using the average entropy $\mathfrak{E} = 3.48$ and the average number of locations visited by the users 62.35 of all users, the average predictability increases to 31%. In both cases, the mobility of the active users is less predictable than 31%.

As stated earlier, in almost 38% of the cases the users move to new locations where they have never been before, but many of these locations have already been visited by friends of the users. Indeed, 74.6% of the 38% new locations have already been visited by one of the friends of the users, thus the amount of unknown locations in the circle of friends of the user reduces to 9.8%. Further, in almost 13% of cases the check-in of a user is followed by a check-in of a friend within one hour, which means in about 13% of the cases the users are involved in a social situation. Two thirds of the active users are already involved in a social situation. The aforementioned statistics show a high potential of at least 10% for improving the prediction accuracy of the mobility model based on social influences.

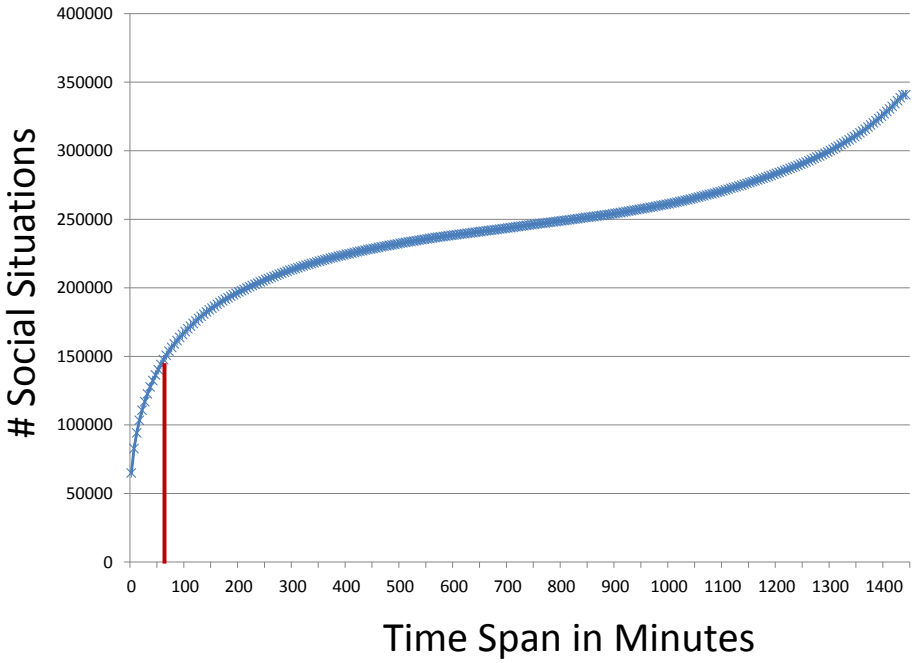

Figure 5.2: The accumulated total number of social situations within a Δt represented by the x axis. For example, setting Δt to one hour, the data set contains less than 150,000 cases where a visit of a user u_i is followed by a visit of a friend u_j within a time span of one hour.

Figure 5.2 plots on the x-axis the time span (in minutes) during which a user follows a friend to a location, and on the y-axis the accumulated number of visits within a time span. The social situations detected by setting Δt to a lower value leads to the detection of less but more realistic social situations. The number of social situations increases when building social situations using a time span Δt greater than an hour, but setting Δt to a higher value leads to the detection of more unrealistic social situations. Thus, the full potential of accuracy improvement due to social influences is higher than 10%.

5.5.2 Performance Analysis

5.5.2.1 Social Influence Estimations

The spatial-temporal (ST) PPM VOMM model from chapter (3) is able to predict the next location of a user with an accuracy of 18.6% by setting the order of the model to two, $\Delta t = 1$ and the average stay time κ to three. The SOST PPM VOMM model is able to incorporate social influences beside spatial and temporal features. Section (5.3.1.3) provides two models for estimating social influences in SOST PPM VOMM according to equations (5.7) and (5.8). The accuracy of SOST PPM VOMM increases to 21.2% when estimating synchronous specific social influences according to equation (5.7), and increases to 22.5% when estimating synchronous specific social influences according to equation (5.8), because estimation according to equation (5.8) gives social influences more importance than equation (5.7) as stated earlier in section (5.3.1.3). The absolute improvements in accuracy corresponds to 0.026 and 0.039, the relative improvements in accuracy corresponds to 14% and 21% respectively. The empirical results using both estimations imply that social influences in general have a high impact on the mobility behavior of human beings.

5.5.2.2 Drift Function

The prediction accuracy of SOST PPM VOMM increases significantly when applying both the drift functions introduced in section (5.3.1.5). The improvement in prediction accuracy is slightly lower, i.e. 0.13% relative improvement in accuracy when using the exponential drift function according to equation (5.13) compared to the power drift function according to equation (5.12). The prediction accuracy of SOST PPM VOMM using equation (5.8) for estimating social influences and setting the degree of drift to $\beta = 0.02$ increases to 23.1%, which corresponds to an absolute improvement in accuracy of 0.006 and a relative improvement in accuracy of 2.7%. The prediction accuracy of SOST PPM VOMM using equation (5.7) for estimating social influences and setting the degree of drift to $\beta = 0.05$ increases to 23.8%, which corresponds to an absolute improvement in accuracy of 0.026 and a relative improvement in accuracy of 12%. The prediction accuracy of SOST PPM VOMM is higher when synchronous specific social influences are estimated according to equation (5.8) than compared to equation (5.7), but contrarily, when applying the drift function the prediction accuracy of SOST PPM VOMM using equation (5.7) for estimating synchronous specific social influences is higher by 3% in comparison with equation (5.8).

Integrating a drift function increases the prediction accuracy of SOST PPM VOMM significantly, regardless of the synchronous specific social influence estimator used. Two two-sided unpaired Students t-tests for both estimators confirmed the significance of the improvement in accuracy ($P(\epsilon) = 2.5{*}10^{-31}$ for the estimator according to equation (5.7) and $P(\epsilon) = 0.02$ for the estimator according to (5.8))

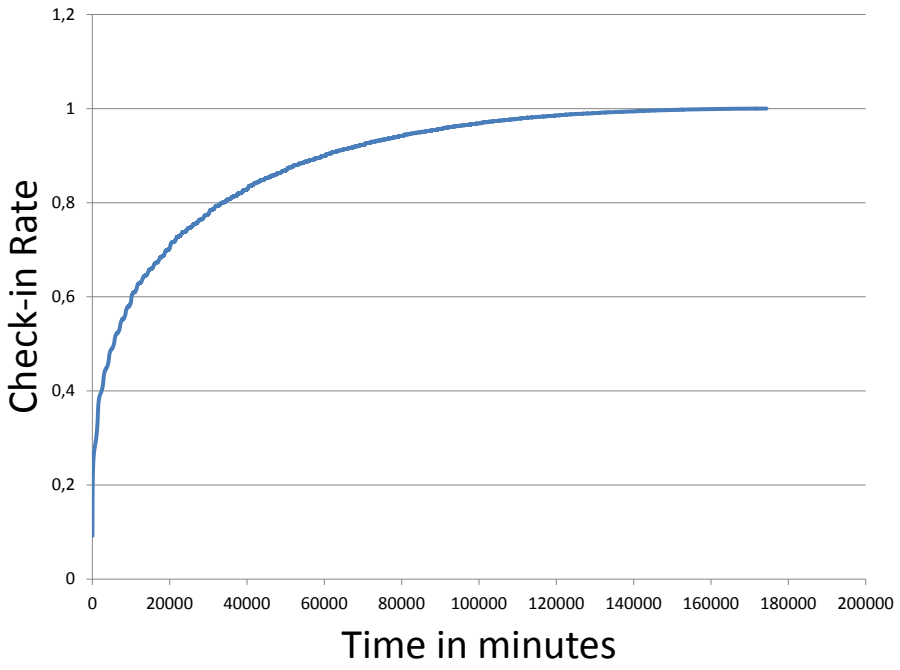

Figure 5.3: The y-axis shows the share of check-ins and the x-axis the time span in minutes for which a visit of a user follows a visit of a friend.

We tried different values for the degree of drift for individual location histories, the maximum prediction accuracy is achieved by setting $\alpha = 0.0005$, which is very low, because the dataset is very sparse and the period of data short. Therefore, we believe that the degree of drift of $\alpha = 0.007$ as stated in chapter (3) is more reliable. The value of $\alpha = 0.007$ implies that the influence of a previous visit of a user to a location on the future mobility of that user decays within two years. The degree of drift given by the hyper parameter β maximizes the prediction accuracy when setting its value to 0.02 and 0.05 for the estimators of synchronous specific social influences according to equations (5.8) and (5.7) respectively. The value of β implies that social influences decay in three to six weeks, which is significantly faster than the decay of individual location history. The value of β is in perfect accordance with figure (5.3), which shows the percentage of cases a user follows a friend to a location within a specified time period (in minutes). Figure (5.3) shows that in almost $\simeq 85\%$ of the cases a user follows a friends to a location within six weeks, and in only 15% of the cases does a user follow a user to a location within a period greater than six weeks.

5.5.2.3 The Impact of Tie Strength

The accuracy of the model increases when using the version of the Jaccard Coefficient that considers the tie strength between the users in equation 5.9 compared to the original Jaccard Coefficient in equation 5.5. Consideration of tie strength in equation (5.9) improves prediction accuracy by 0.002 which corresponds to 0.8% relative improvement in accuracy compared to the pure Jaccard coefficient of equation (5.5). A two-sided unpaired Student's t-test for checking the significance of the improvement in accuracy shows only minor significance (0.18).

Equation 5.10 calculates the tie strength between two users based on the co-locations (spatial overlap) in the location histories of the two users. Equation 5.11 weights the co-locations according to the importance of the location for the user, for whom the next location is going to be predicted (tie strength is not transitive, see section (5.3.1.4)). We used different weighting factors from section 4.4.3. None of the weighting factors could improve the accuracy significantly, possibly because of the sparsity of the location history of the users (a check-in in Foursquare is voluntarily, the users make check-ins sporadically at locations they find interesting and avoid check-ins at locations where privacy is an issue).

5.5.2.4 Synchronous Specific Social Influence

Table (5.2) shows the improvement in accuracy achieved by incorporating synchronous specific social influence factors compared to the accuracy of the spatial-temporal PPM VOMM. The best results are achieved by setting the hyper-parameters to $\beta = 0.05$, $\kappa = 3$ and $\Delta t = 1$.

s-class	Abs. Improvement	Rel. Improvement %	Two-Sided Unpaired T-Test $P(\epsilon)$
Class I:	0.0088 (0.0118)	4.7 (6.1)	0.0012 (0.00113)
Class I & II:	0.0235 (0.0313)	12.6 (16.3)	$3.8 * 10^{-28}$ ($1.4 * 10^{-43}$)
Class I-III	0.0344 (0.0458)	18.5 (23.9)	$1.4 * 10^{-39}$ ($1.5 * 10^{-76}$)

Table 5.2: Empirical results: Column 2 represents the absolute improvement in accuracy compared to ST PPM VOMM model, column 3 represents the relative improvement in accuracy compared to ST PPM VOMM model, column 4 the results of two-sided unpaired t-tests (probability of error $p(\epsilon)$) for showing the significance of the improvements. The numbers in braces represent the corresponding values for the portion of users who are involved in at least one social situation (setting $\beta = 0.05$ and $\Delta t = 1$, $\kappa = 3$).

The accuracy of the spatial-temporal model is 18.6%. Inclusion of class I synchronous specific social influence factors increases the prediction accuracy to 19.48% for all users and for the users who are at least involved in one social situation to 19.78%. The absolute improvement in accuracy is 0.88 and 1.18, the relative improvement in accuracy is 4.7% and 6.1% respectively.

Incorporation of class II synchronous specific social influence factors causes the accuracy to increase to 20.95% for all users and to 21.73% for the users with at least one social situation. The accuracy improves to 22.04% when class III synchronous specific social influence factors are also incorporated. The total absolute improvement in accuracy by incorporating all classes of synchronous specific social influence

factors is 0.0344, the relative improvement in accuracy corresponds to 18.5%. Inclusion of class II & III social influence factors injects a vast amount of knowledge to the individual mobility model of a user. The empirical results underline the importance of social influences for improving the accuracy of next location prediction. The improvement in accuracy in all cases is statistically significant, corresponding Student's t-tests shown in table (5.2) confirm the significance of the results.

5.5.2.5 General Social (Trends) Influences

Almost one third of the users were not involved in any social situation, and an improvement in accuracy can only be achieved by incorporating general (trends) social influences. The numbers in braces in table (5.2) show the empirical results for those users who are involved in a minimum of one social situation.

General social trends/influences have a significant impact on the mobility of a user. General social influences occur even a user is not involved in an obvious social situation. General social influences occur either due to trends in the community of a user or due to transferring social influence via other media such as internet, phone, etc. Thus, a user might be under social influence even if they move alone. Many people recommend restaurants, shopping malls, sport activities, cinema films, art exhibitions, etc. to their friends via phone, e-mail, SMS, online social networking platforms, etc. and these friends follow them with a time delay to these locations.

The incorporation of general social influence factors improves the accuracy of prediction further to 23.8%. The additional absolute improvement in accuracy due to the incorporation of general social influences is 0.0178 which corresponds to a relative improvement in accuracy of 8%. The total improvement in accuracy due to the incorporation of social influences increases to 0.0522, which corresponds to a relative improvement in accuracy of 28%. The significance of the improvement is confirmed by a two-sided unpaired Student's t-test ($P(\epsilon) = 2.0 * 10^{-19}$), the significance of the total improvement in accuracy is confirmed by a further two-sided unpaired t-test $P(\epsilon) = 2.6 * 10^{-102}$. The importance of general social influences is emphasized by considering only users who are not involved in any social situation. Almost one third of the users are not involved in any social situation. Incorporating synchronous specific social influence factors does not lead to an improvement in accuracy for those users. In contrast, the incorporation of general social influences leads to an absolute improvement in accuracy of 0.0221, which corresponds to a relative improvement in accuracy of 11%. This result shows that even if a user is not accompanied by friends, they might still be under social influence.

5.5.2.6 Assessment of the Improvement in Accuracy

The total accuracy when considering all kinds of social influence factors improves to 23.8%. The total absolute improvement in accuracy when considering all kinds of social influences increases to 0.0522, which corresponds to a relative improvement in accuracy of 28%. The significance of the improvement in accuracy can be emphasized by a two-sided unpaired Student's t-test ($P(\epsilon) = 2.6 * 10^{-102}$) confirming

the importance of social influence factors for improving individual next location prediction.

The accuracy of the SOST PPM VOMM predictor is low because the users are high entropic, thus their mobility is less predictable. According to Fano's inequality [Fano, 1961], on average, the users are predictable to 29% as shown in section (5.5.1). The prediction power of SOST PPM VOMM can be demonstrated taking the predicability of the users into account. SOST PPM VOMM is able to achieve a prediction accuracy of 23.8% from a maximum predictability of 29% , which corresponds to an accuracy of at least $23.8/29 > 82\%$.

The impact of social influences on improving location prediction can be convincingly demonstrated, if only those users who were involved in a minimum of one social situations are considered. For those users, the total absolute improvement in accuracy increases to approximately ≈ 0.0615 which corresponds to a relative improvement in accuracy of 32%. The absolute accuracy improvement for users involved in at least 10 social situations actually rises to ≈ 0.085 which corresponds to a relative improvement in accuracy of 43%.

5.5.2.7 The Prediction of Unknown Locations

The lack of sufficient location history leads to the zero-frequency problem. This problem occurs when an observation occurs for the first time during the test phase. As stated earlier, incorporating the location history of friends into the model adds a vast amount of knowledge to the individual mobility model and enables locations to be predicted which have been visited by friends, but where the user has never been before. Integration of social networks resulted in being able to make a total number of 437 231 non-trivial (non-trivial == without having to refrain to the uniform escape probability) predictions of the next location, when the current location was visited for the first time ever. Of these 437 231 predictions, 36 469 were successful. The absolute improvement in total prediction accuracy by incorporating these non-trivial 437 231 predictions was 0.0319 which corresponds to 61% of the total absolute improvement in accuracy. The significance of the improvement is confirmed by a two-sided unpaired Student's t-test ($P(\epsilon) = 8{*}10^{-11}$). The best improvement in accuracy for predicting unknown locations was during the hours from seven p.m. to 12 p.m. on work days and after 11 a.m. on weekend days when people usually spend time on social activities. Figure (4.1) from the previous chapter shows that most of the check-ins are indeed at public locations in the categories that are tightly related to social activities such as eating, entertainment, nightlife, outdoor activities and shopping. The prediction of unknown locations once again emphasizes the importance of social networks in next location prediction.

5.5.2.8 The Distribution of Improvements in Accuracy

Humans tend to have free time during the evening hours of all days or during the afternoon hours of weekend days. Examples of free time activities are a visit to a swimming pool, museum, theatre, cinema, sport stadium, music concert, fitness studio, etc. Further examples of free time activities are visits to a restaurant, a

favorite bar, a shopping mall, etc. Most of the above activities simultaneously are social activities, which means people are usually accompanied by their friends during these activities. As stated earlier, figure (4.1) from the previous chapter confirms that indeed most of the check-ins are at public locations typical for social activities.

Further, people are most explorative during the aforementioned time periods, thus the uncertainty of predicting their next location is the highest. Therefore, we believe that specifically during these hours information gleaned from the social network can contribute to reducing the uncertainty of prediction.

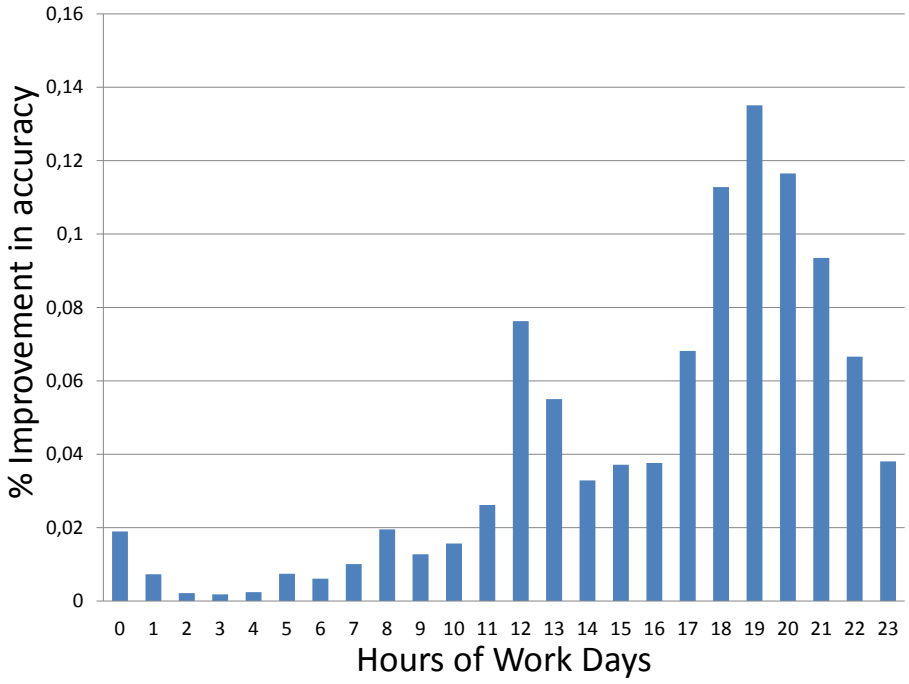

Figure 5.4: The proportion of absolute improvement in accuracy over a work day.

Figure (5.4) shows that the improvement in accuracy on work days has two peaks. The first peak occurs at lunch time when it is likely that people meet their friends for lunch. The second peak is during the evening hours when people meet friends after work for a drink in a bar, for dinner in a restaurant or for a window shopping stroll around the city or shopping Mall, etc.

Figure (5.5) plots the proportion of improvement in accuracy for the hours of the weekend days. The plot shows a maximum plateau rather than peaks, which covers a period between 11 a.m. and 12 p.m, which is in accordance with the typical behavior of humans. Humans clearly spend time with their friends between 11 a.m. and 12 p.m. during the weekend.

Both figures show that the most improvement in accuracy due to social influences occurs during the time periods with the highest uncertainty for next location prediction. The figures impressively underline the importance of social influences for improving the prediction accuracy during time periods with high uncertainty, when the individual mobility model fails to find mobility patterns.

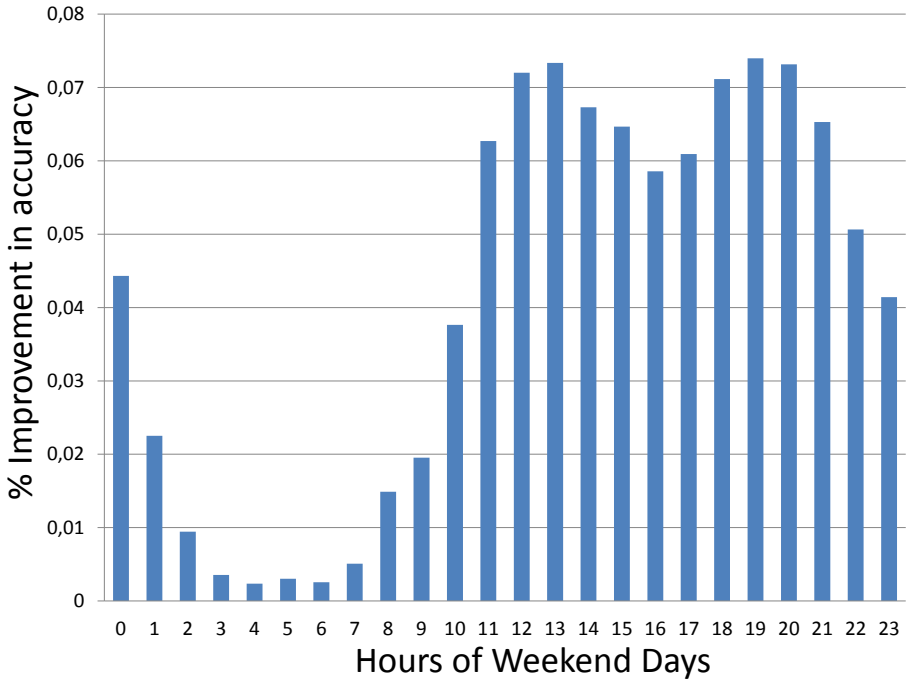

Figure 5.5: The proportion of absolute improvement in accuracy over weekend days.

5.5.3 Social Network Measurements

This subsection focuses on the empirical analysis of the correlation between various social network measurements and improvement in accuracy from including the location histories of friends.

5.5.3.1 Number of Influencers

An influencer is a user whose mobility influences the mobility of a friend who we denote as follower. A user (potential follower) visits a location q at time step t after previous visits of friends (potential influencers) to the same location q at earlier time steps $< t$. The majority of users (potential followers) in the Foursquare dataset are influenced by a few friends (potential influencers). Fewer than 1% of the users have more than 20 potential influencers. Figure (5.6) shows the relationship between potential influencers (a-axis) and potential followers (y-axis) using a logarithmic scale for both axes. The plot shows a power law relationship between both followers and influencers with a coefficient of determination of 0.97, confirming that the most of the users follow only a few potential influencers.

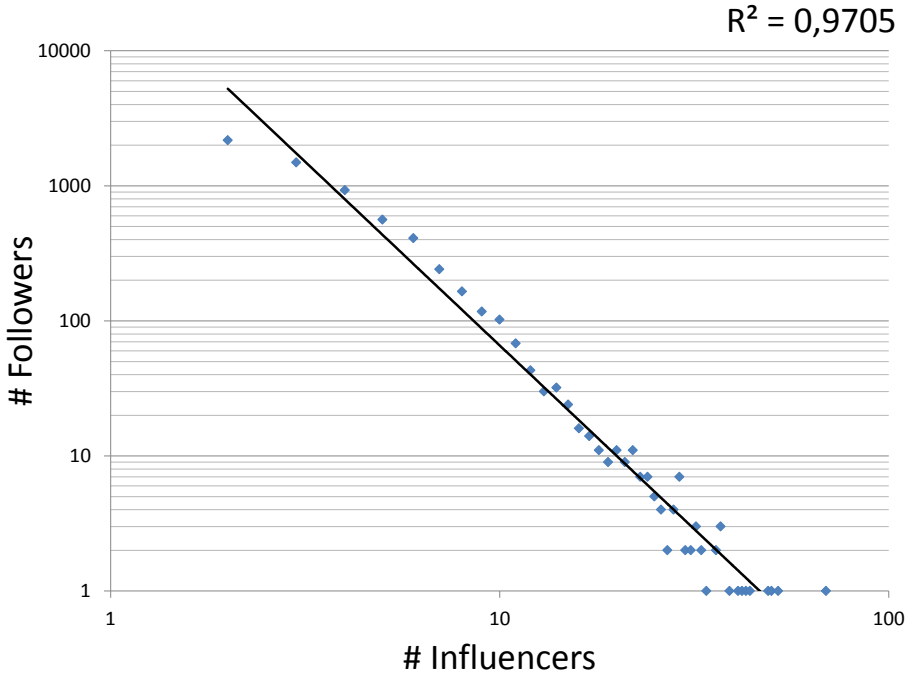

Figure 5.6: The log-log relationship between followers (y-axis) and influencers (x-axis).

Inclusion of the location histories of a few friends (influencers) is sufficient to improve prediction accuracy significantly. The inclusion of the location histories of only two friends is sufficient to improve prediction accuracy of an individual by 0.0415 absolute improvement and 21% relative improvement compared to the mobility model based on the location history of the individual only. The inclusion of location histories of more friends increases the improvement in accuracy more. Figure (5.7) plots the relationship between the number of influencers (x-axis) and the improvement in accuracy. The plot shows a positive trend, which is confirmed by a moderate positive correlation coefficient according to Pearson's correlation coefficient ($r = 0.37$) and a strong positive correlation according to Spearman's rank correlation coefficient ($\rho = 49$) with a probability of error of zero ($\epsilon = 0.0$).

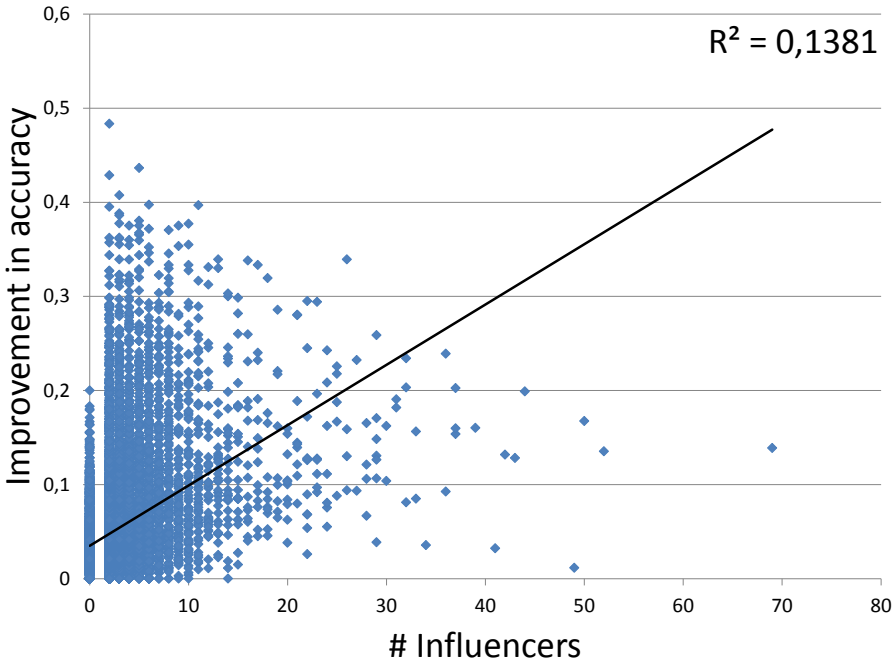

Figure 5.7: The relationship between the number of influencers (y-axis) and absolute improvement in accuracy (x-axis).

5.5.3.2 Injected History Size

The size of location histories injected by the influencers exhibits the same trend as the number of influencers (both general and synchronous specific social influencers). The inclusion of only 50 visits by influencers is sufficient to improve the accuracy by a significant 0.0266 absolute improvement and 14% relative improvement. Figure (5.8) plots the average size of the location history of influencers which has been injected. The plot shows a positive trend showing that the inclusion of more location history from influencers improves accuracy. The positive trend is confirmed by a moderate positive correlation of 0.23 according to Pearson's correlation coefficient, and a similar positive correlation of 0.21 according to Spearman's correlation coefficient with a probability of error of zero ($\epsilon = 0.0$).

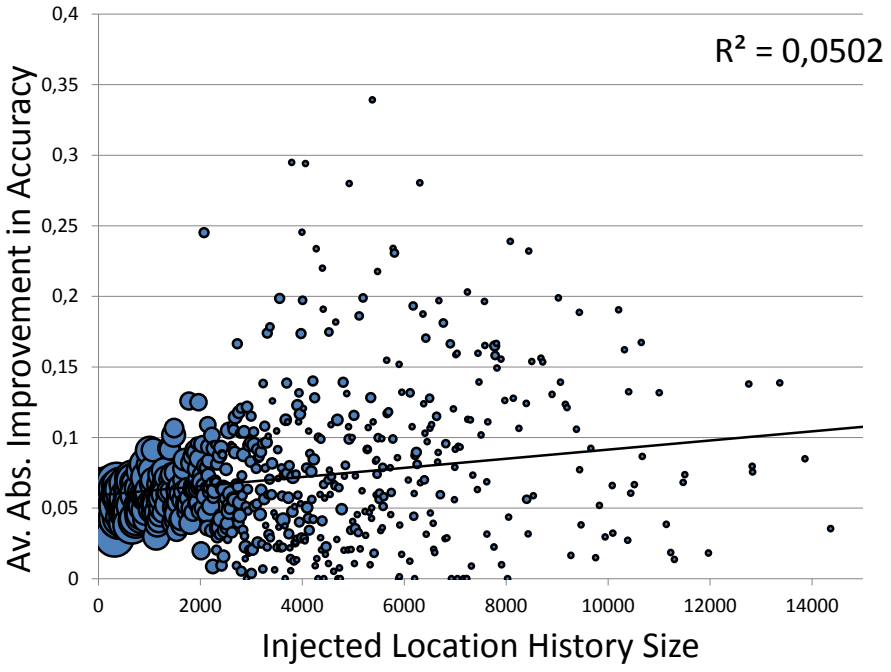

Figure 5.8: The relationship between absolute improvement in accuracy and the size of the injected location history from influencers exhibits a positive correlation with correlation coefficients $r = 0.23, \rho = 0.21, P(\epsilon) = 0.0$. The size of the bubbles indicates the number of users at that data point.

5.5.3.3 Social Situation Rate

Synchronous specific social influence factors are integrated in SOST PPM VOMM using social situations, thus it is expected that the number of social situations correlates positively with improvement in accuracy. Figure 5.9 shows a positive trend between the improvement in accuracy and the social situation rate, which means a higher social situation rate implies a higher improvement in accuracy. The positive trend is confirmed by both Pearson's and Spearman's correlation coefficients. The correlation coefficients were found to be 0.71 and 0.61 respectively. Users with higher social situation rates represent extrovert users, who are socially active and thus their mobility behavior is more predictable via social amendments to the spatial-temporal ST PPM VOMM approach (due to synchronous specific social influence factors). Users with lower social situation rates may be considered as more introvert users and their prediction accuracy can be improved less by using synchronous specific social influence factors, but nonetheless their prediction accuracy may still be improved using general social trends.

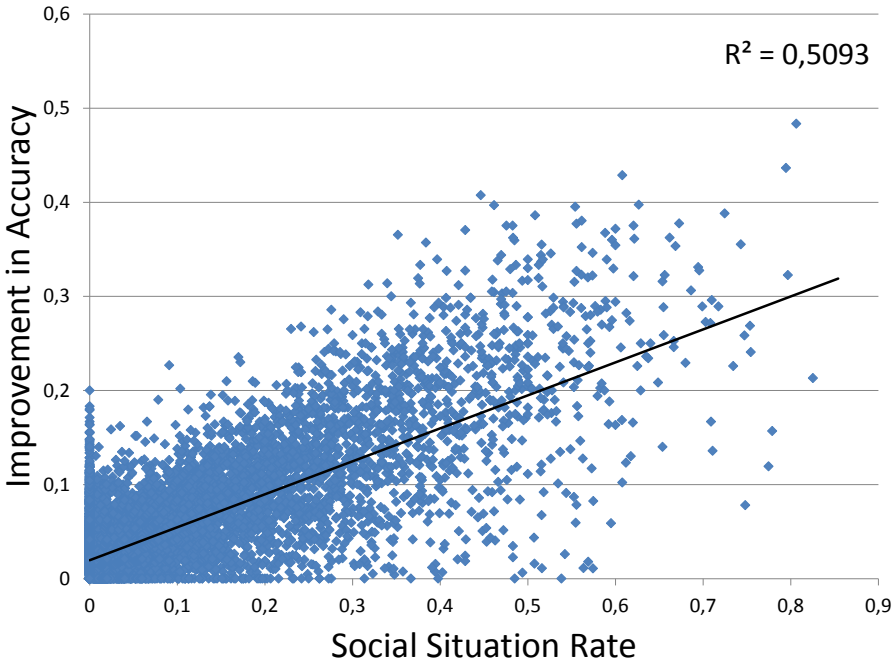

Figure 5.9: Absolute accuracy improvement correlates with the average social situation rate $r = 0.71, \rho = 0.61, P(\epsilon) = 0.0$.

5.5.3.4 Cohesive Subgroups (Strong Ties)

The dataset contains 149 700 social situations that have been integrated in the SOST PMM VOMM tree. As many as 105 087 of the social situations are between the members of the same cohesive subgroups (maximal 2-plexes), which corresponds to more than 70% of the social situations. Accordingly, most of the improvement in accuracy is due to social situations among members of the same maximal 2-plexes. This result emphasizes the importance of cohesive subgroups for transferring influence in the circle of friends and is in accordance with the results of the correlation analysis between the measure of cohesion in maximal 2-plexes and mobile homophily in section (4.6.2.2).

The number of users involved in a social situation defines its size. The size of a social situation varies between 2 and 16 users. Figure (5.10) plots the size of social situations (x-axis) and the number of social situations (y-axis). The distribution of number of social situations over their size follows a power law with a coefficient of determination of $R^2 > 0.99$, which means most of the social situations are of a size between two and five, and only a small portion are of a higher size.

The measure of cohesion according to equation (4.2) can be directly applied to the users involved in a social situation, but since we are interested in investigating the influence of maximal 2-plexes on the improvement of next location prediction, and since the measure of cohesion between the users involved in a social situation does not necessarily reflect the true cohesion between the user and the friends involved

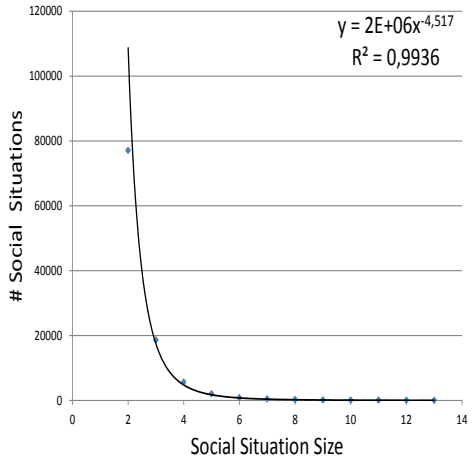

Figure 5.10: A comparison of the distribution of absolute accuracy improvement over the hours of a work day for all social situations (red bars) and social situations between the members of the same 2-plex (blue bars).

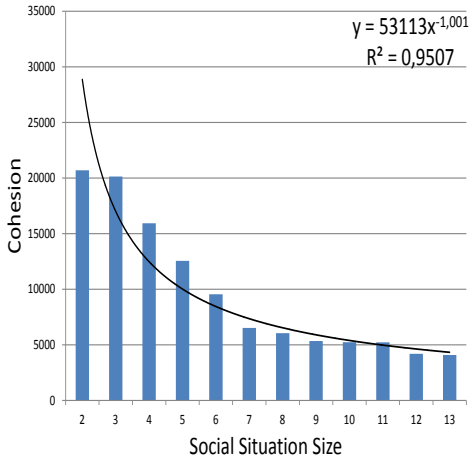

Figure 5.11: A comparison of the distribution of absolute accuracy improvement over the hours of weekend days for all social situations (red bars) and social situations between the members of the same 2-plex (blue bars).

in the social situation, because the social situation may contain friends from the same maximal 2-plexes and others from outside. Further, many social situations are only between two users which does not allow a reliable calculation of cohesion. Because of the aforementioned reasons, the cohesion inside the whole social situation is lower compared to the cohesion between the user in question for prediction and another user in the social situation if the two users are members of the same 2-plex, therefore, we calculated a measure of cohesion for each social situation as follows. We identify a set $\mathfrak{P}$ of all maximal 2-plexes which contain at least the user and one of their friends involved in the social situation. We calculate the measure of cohesion of the social situation by averaging the measure of cohesion of all maximal 2-plexes in $\mathfrak{P}$. Figure (5.11) plots the measure of cohesion of the social situations (y-axis) against their sizes (x-axis). The distribution follows a power law with a coefficient of determination of $R^2 > 95$. The higher the number of users involved in the social situation the lower the measure of cohesion, which is in perfect accordance with the measure of cohesion of the maximal 2-plexes in the previous chapter (figure 4.4).

The greatest improvement in accuracy should thus be achieved with social situations of a lower size if we assume that cohesion and prediction accuracy are positively correlated. Figure (5.12) sets the proportion of improvement in accuracy and the size of social situations in relation to one another. The relationship follows a power law with a coefficient of determination of $R^2 > 99$, which is in accordance with the distribution of social situations over their sizes. Figure (5.13) sets the average measure of cohesion for the different social situation sizes in relation to the proportion of improvement in accuracy. Similarly to the relationship of the measure of cohesion and the size of social situations, the distribution follows a power law with a coefficient of determination of $R^2 > 95$. In accordance with the correlation analysis in the previous chapter, a strong positive correlation exits between accuracy improvement and the measure of cohesion.

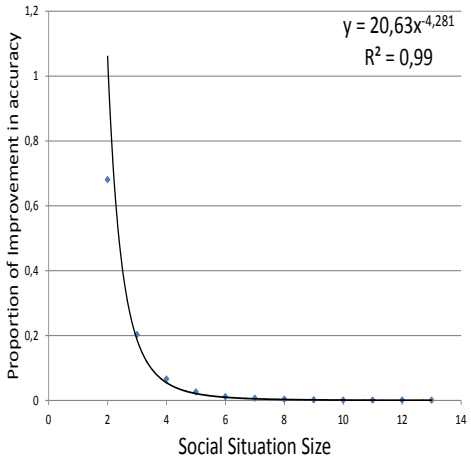

Figure 5.12: The relationship between the percentage of total absolute improvement in accuracy and the size of social situations follows a power law with a coefficient of determination of 0.99.

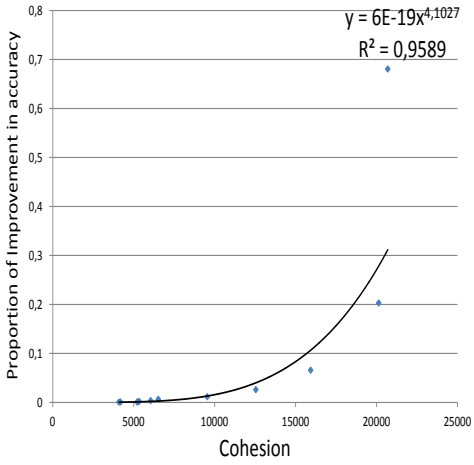

Figure 5.13: The relationship between the percentage of total absolute improvement in accuracy and the average measure of cohesion in the social situations follows a power law with a coefficient of determination of $\backsimeq 0.96$.

5.5.3.5 Degree Centrality & Weak Ties

A user u in a social network has typically many social ties to other users in the social network, which we have already denoted as neighbors. People, typically, maintain their social relationships on different scales, because they have cognitive, emotional, spatial and temporal limits, which prevent them from maintaining all their relationship with the same intensity [Granovetter, 2005]. Dunbar suggests the number of neighbors, with whom a user can maintain stable cognitive social relationships to be 150 ([Dunbar, 1992] as cited by [Dun, 2013]. Almost 9% of the users in the Foursquare dataset have more than 150 neighbors, and 22 of the users have at least 1 000 neighbors, which means that the users have at lot of weak ties. A user with 1 000 neighbors has intuitively more ties, than they can cognitively maintain or interact with. Such a users is most probably a prominent user with a lot of followers (fans), the only interaction between them and their followers occurs via sharing check-ins. The followers are probably keen to visit the same locations, which these prominent users find interesting. A prominent user is thus important for setting/transmitting general social trends.

The neighbors of a user u and their ties contains dense regions of knit clump, most of whom are in touch with one another [Granovetter, 2005]. The users inside such a knit clump build a cohesive subgroup with strong ties among them. We have already found out in the previous chapter, that the majority of maximal 2-plexes are of small size. The biggest maximal 2-plex found was containing 20 users, a user can maintain tight social relationships to subset of their friends less than 20 friends, which is in perfect agreement with the human social perception limit in [Fischer and Wiswede, 1997]. The sizes of maximal 2-plexes again emphasize that most of the social relationships in the Foursquare dataset represent weak ties. The focus of the previous section was on strong ties, in this section we investigate the importance of

weak ties in social networks.

The neighbors of a user can be subdivided into strong and weak ties. Each user in a social network have a knit clump of strong ties, and a lot of weak ties to less close friends, let us say acquaintances (in accordance to Granovetter). Each acquaintance is in turn enmeshed in their own (different) knit clump of strong ties, thus a weak tie is a bridge between two knit clumps of social ties. The weak ties in a social network are responsible for connecting different communities and social circles together [Granovetter, 1982]. Although the behavior of a typical user in a social network is rather similar with the behavior of their close friends, but the weak ties are important for transmitting general social trends beyond the borders of the own social group. Users interact and spend time mostly with their close friends, thus they exhibit some kind of homogeneity which results in a considerably high overlap in their behavior, life styles, emotional needs, thoughts, beliefs, movements, goals, information, etc. Weak ties provide users with more novel information because of the heterogeneity in their beliefs, thoughts, goals, information, emotional needs, etc. The heterogeneity in the information of two users connected via a weak tie occurs because each of the two users spends time and interacts with people, who the other user does not know and, thus, the two users exchange rather more novel information [Granovetter, 2005].

Degree centrality is a notion that refers to the extent in which a user is connected to others. A user with a lot of friends is more central or prominent and is important for the flow of information in the social network, because they have a lot of weak ties and get in touch with numerous others from different social groups. As shown by [Gladwell, 2002], there are people in every social group whose social circle is four or five times the size of other people's. These people have the habit of making introductions. Gladwell refers to these people as connectors, we prefer to call them central users. Central users are a few people who have the extraordinary knack of making friends and acquaintances and who can bring users from different social circles together ([Gladwell, 2002, Pages 38-41] as cited by [tip, 2013]). The best explanation for the abilities of connectors is probably Gladwell's, "Their ability to span many different worlds is a function of something intrinsic to their personality, some combination of curiosity, self-confidence, sociability, and energy." ([Gladwell, 2002, Page 49] as cited by [tip, 2013]).

As stated earlier, the foursquare dataset contains 22 users with at least 1000 friends, we refer to these users with central users (in the sense of degree centrality). We enclosed the users with more than 1 000 in red ellipses in the following three figures. The Foursquare dataset contains 147,900 social situations, almost $> 70\%$ of the social situations are between members of the same maximal 2-plexes, i.e. between users connected with a strong tie. The social situations between the strong ties are responsible for 88% of the total improvement in accuracy. The remaining 30% of social situations are responsible for 12% of the improvement in accuracy. It can be easily seen that the users spend most of their time with their close friends, but nevertheless, the users interact in 30% of the cases with their acquaintances.

Pursuing the idea of degree centrality, figure 5.14 represents the relationship between the average degree and the improvement in accuracy. The focus of this section is rather on the existence of the central users than on a trend between improvement

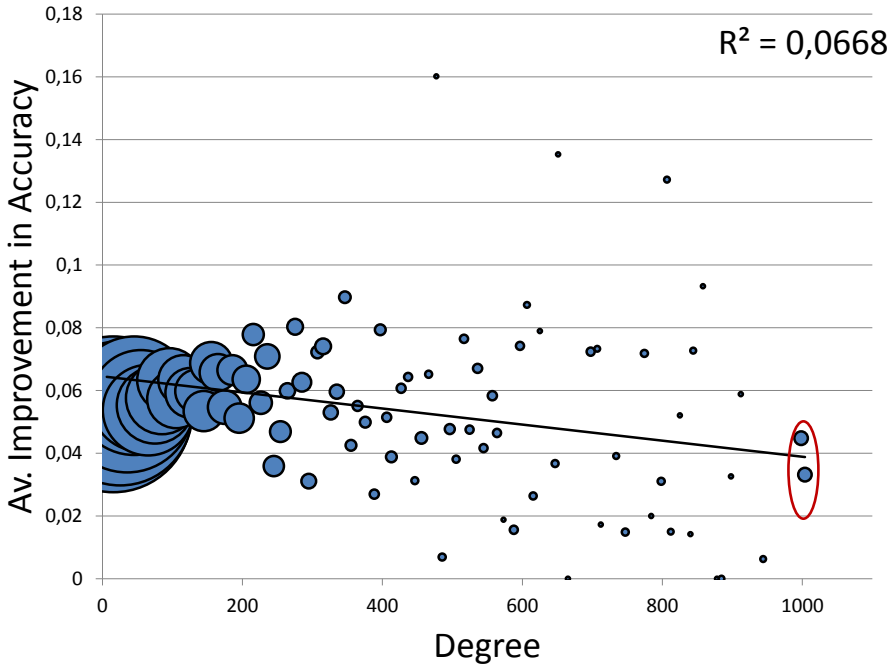

Figure 5.14: The average absolute improvement in accuracy shows a negative trend as the degree increases, the correlation coefficients were found to be ($r = -0.26, \rho = -0.29, P(\epsilon) = 0.0$).

in accuracy and average degree of the users, nevertheless, a negative trend is easy to see between both quantities. The negative trend means that the mobility of central users (red ellipse) is only slightly improved using the location histories of their friends, because a central user interacts more with their weak ties from different social communities. The low improvement in accuracy for the central users is in accordance with their extraordinary ability to span different social circles as stated earlier. A user who has 1 000 neighbors can influence the mobility of much more users than a user with few neighbors. For example, a check-in of a central user at a location in Foursquare can be seen by 1 000 neighbors. A part of the neighbors may visit the same location during a later time period, Influenced by this check-in. We conclude that central users are trend setters or trend transmitters between different social communities and are followed by rather than following others.

Central users get about a lot and are explorative in nature. Figure (5.15) plots the relationship between the degree and the average size of location history. The plot shows a positive trend $r = 0.16, \rho = 0.15, P(\epsilon) = 0.00016$, confirming that central users do get about a lot. Figure (5.16) plots the relationship between the degree and the average number of locations visited for the first time. The plot shows a positive correlation $r = 0.33, \rho = 0.25, P(\epsilon) = 0.0$ confirming the explorative nature of central users. We conclude from the above results that although the prediction accuracy of central users is not improved very much by including the location histories of their friends, but they are very important, because their mobility contributes greatly to improving the prediction accuracy of many other users.

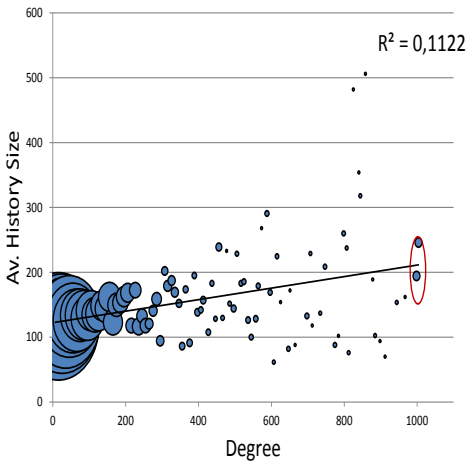

Figure 5.15: A plot showing the positive trend between the degree and the average location history ($r = 0.33, \rho = 0.25, P(\epsilon) = 0.02$). The size of the bubbles indicates to the number of users with a given degree.

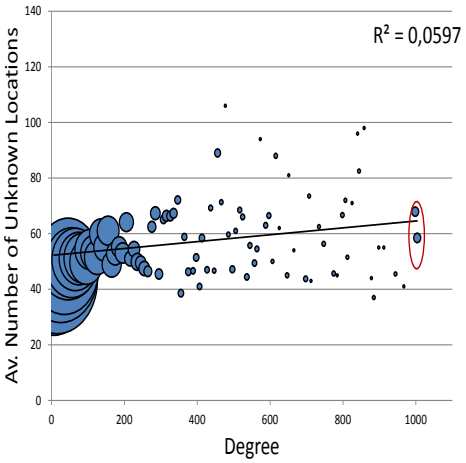

Figure 5.16: A plot showing the positive correlation between the degree and the average number of locations visited for the first time ($r = 0.24, \rho = 0.21, P(\epsilon) = 0.05$). The size of the bubbles indicates to the number of users with a given degree.

5.5.4 Location History Measurements

The focus of this subsection is on investigating the correlation between the improvement in accuracy and various measurements relating to the location history of the user such as average location history per location, entropy, the number of visited locations per user, etc. in order to emphasize the importance of social influences for enhancing prediction accuracy.

5.5.4.1 History Size

The size of the available training sequence has a major influence on the performance of all predictive models based on probabilistic reasoning. As has already been stated in previous chapters, the performance of such models depends to a high extent on the availability of sufficient training data per state (location). It would seem likely that by including social networks, accuracy would be improved only for those users with insufficient training data, in other words a sufficiently large training sequence makes the inclusion of location histories of friends obsolete (apart from visits to new locations, that are not yet contained in the location history).

Figure (5.17) sets both the size of location history and the improvement in accuracy in relation to one another. The figure shows only a very slight negative trend $r = -0.06, \rho = -0.06, P(\epsilon) = 0.0$. The result seems surprising, the improvement in accuracy due to the inclusion of social influences seems to be less dependent on the size of training data. A closer look at the location histories of the users provides a plausible explanation as follows:

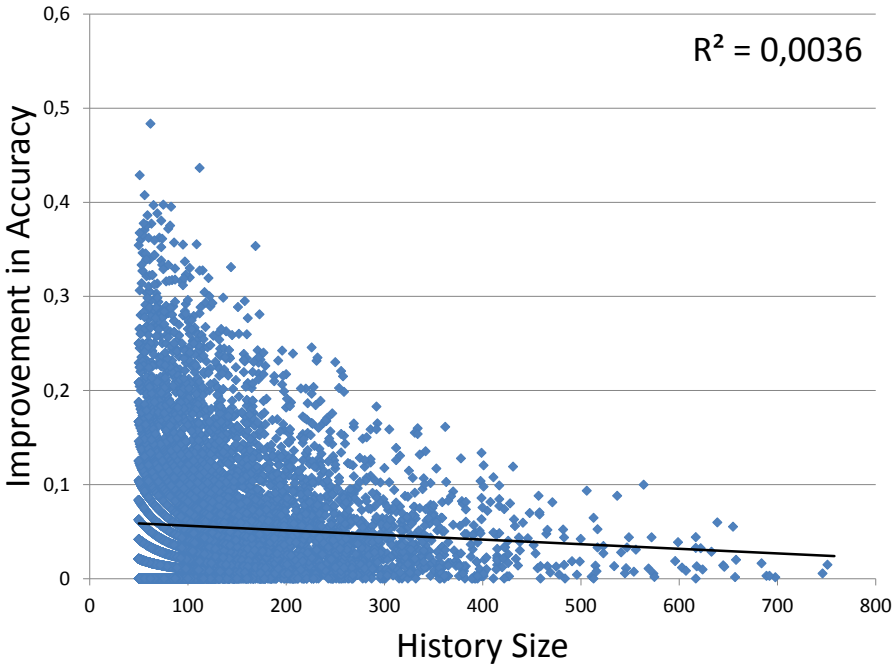

Figure 5.17: The absolute improvement in accuracy shows only a slight negative trend in relation to the size of training data by including social network influences $r = -0.06, \rho = -0.06, P(\epsilon) = 0.0$.

As stated earlier, human beings are explorative in nature and love to visit new and interesting locations. Further, the mobility behavior of users is subject to huge change over the course of years, each user has only a few significant locations such as "home" and "work" and a lot of less significant locations, which are visited irregularly. This means a long sequence of location history contains more locations and causes the entropy of the user to increase. Figure (5.18) plots the size of location history (x-axis) and the number of locations contained in the location history. The plot shows a very strong positive trend, the number of locations correlates very strongly positive to the size of the training data. The strong positive correlation is confirmed by both Pearson's and Spearman's correlation coefficients ($r = 0.77, \rho = 0.75, P(\epsilon) = 0.0$ respectively). Figure (5.19) sets the size of location history in relation with average number of locations, the trend becomes even more obvious, both correlation coefficients increase to $r = 0.79$ and $\rho = 0.85$.

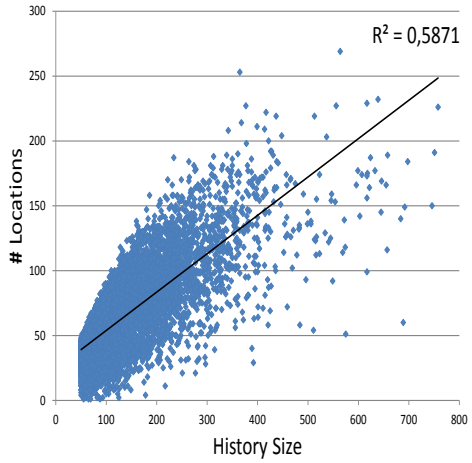

Figure 5.18: The number of locations visited by a user increases as the size of their location history increases $r = 0.77, \rho = 0.75, P(\epsilon) = 0.0$

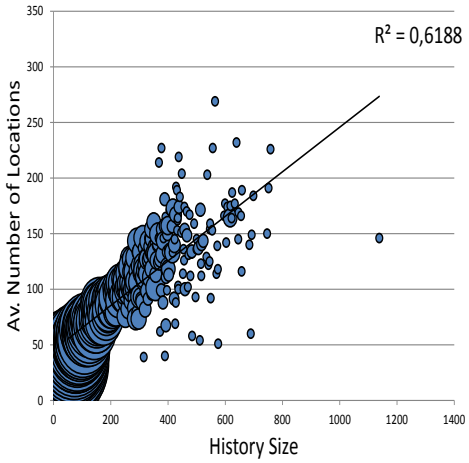

Figure 5.19: The number of locations visited by a user increases as the average size of their location history increases $r = 0.79, \rho = 0.85, P(\epsilon) = 0.0$.

Figure (5.20) sets the size of training data and user entropy in relation to one another. Similarly to the number of locations, the entropy shows a positive trend compared with the size of location history. The users in the Foursquare dataset visit more and more locations with similar frequency of visits as time goes on. This trend is confirmed by the positive correlation between entropy and history size according to Pearson's $r = 0.26$ as well as Spearman's $\rho = 0.30, P(\epsilon) = 0.0$ correlation coefficients. The correlation between both quantities increases significantly when Setting the average entropy in relation to the size of location history, both correlation coefficients increase to $r = 0.45, \rho = 0.55, P(\epsilon) = 0.0$ showing a strong positive correlation (figure (5.21)).

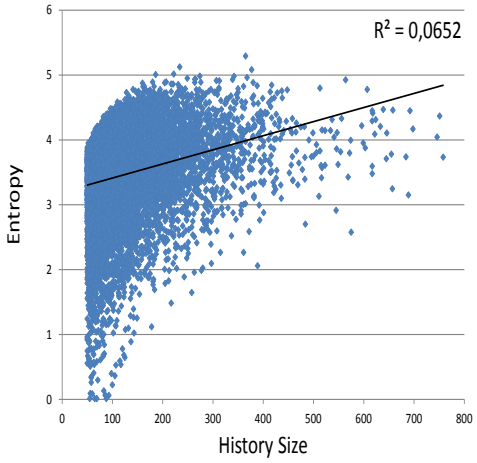

Figure 5.20: The entropy of a user shows a positive trend with the size of their location history $r = 0.26, \rho = 0.30, P(\epsilon) = 0.0$.

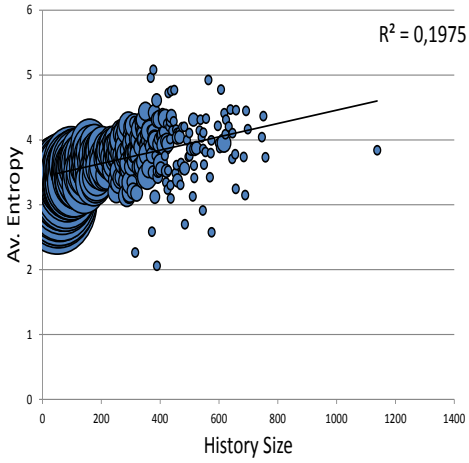

Figure 5.21: The entropy of a user shows a positive trend with the size of their location history $r = 0.45, \rho = 0.55, P(\epsilon) = 0.0$.

The assumption that a larger location history contributes to improving accuracy is only partly true. From the previous explanations it is more than reasonable to conclude that history size only partially contributes to improving accuracy and that consequently, social influences will always remain indispensable for enhancing prediction accuracy.

5.5.4.2 Number of Locations

The average number of locations visited by a user is another quantity that negatively influences the accuracy of a location predictor. Users who are explorative in nature visit a lot of locations with similar probabilities, thus their mobility is less predictable. Incorporating a social network injects a vast amount of data generated by the friends of a user which can be used in training the model and which indeed increase the predictability of the mobility of the user. Figure 5.22 plots the relationship between the average number of locations visited by the different users and the improvement in accuracy. The plot shows a positive trend, which confirms that the inclusion of social networks indeed increases the predictability of explorative users who visit a lot of locations. The positive trend is confirmed by a moderate positive correlation coefficient according to Pearson $r = 0.20$. Spearman's rank correlation coefficient shows an even stronger positive correlation of $\rho = 0.35, p(\epsilon) = 0$ respectively). The trend becomes more clear when setting average improvement in accuracy in relation to number of locations (figure (5.23)). The correlation between both quantities increases, both Pearson's and Spearman's correlation coefficients show a strong positive trend ($r = 0.51$ and $\rho = 0.42$ respectively).

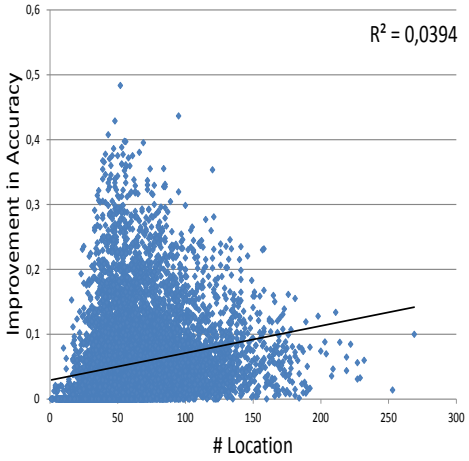

Figure 5.22: The plot shows a positive correlation between the number of locations visited by each user and absolute improvement in accuracy $r = 0.20, \rho = 0.35, P(\epsilon) = 0.0$.

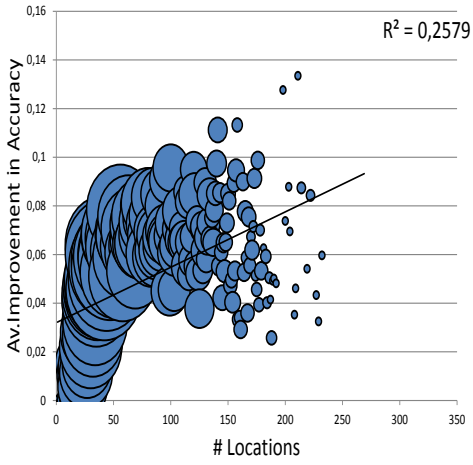

Figure 5.23: The plot shows a positive correlation between the number of locations visited by each user and average absolute improvement in accuracy $r = 0.51, \rho = 0.42, P(\epsilon) = 0.0$.

5.5.4.3 Average Frequency of Visit per Location

Users with a high average frequency of visit per location are the easiest to predict. The average frequency of visit per location in the Foursquare dataset is very low, each user makes on average 2.04 check-ins per location. The predictability of the mobility of users with low average frequency of visit per location is accordingly low. The inclusion of location histories of friends helps increase their predictability. Figure (5.24) plots the average frequency of visit per location for each user with their corresponding improvement in accuracy. The plot does indeed show that including social networks has a stronger positive impact on increasing the predictability of the mobility of users with a low average frequency of visit per location compared to those with a higher average frequency of visit per location. The positive impact is confirmed by a moderate negative correlation coefficient of $r = -0.21$ according to Pearson, and a strong negative correlation coefficient of $\rho = -0.40, P(\epsilon) = 0$ according to Spearman. Figure (5.25) sets the average improvement in accuracy to frequency of visit per location in relation. The negative trend becomes dramatically more clear. A very strong negative trend is easy to see, the trend is confirmed by a strong correlation $r = -0.34$ according to Pearson and a very strong correlation $\rho = -0.80$ according to Spearman.

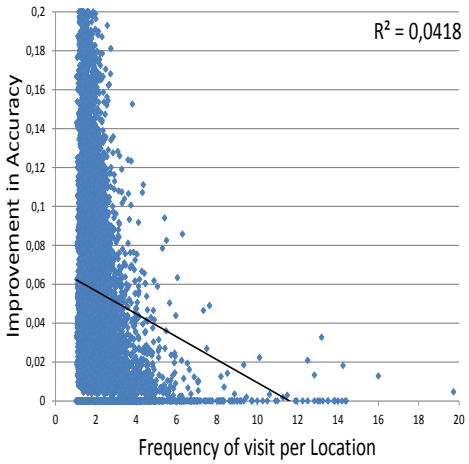

Figure 5.24: The plot shows a negative correlation between the frequency of visit per location and absolute improvement in accuracy $r = -0.21, \rho = -0.40, P(\epsilon) = 0.0$.

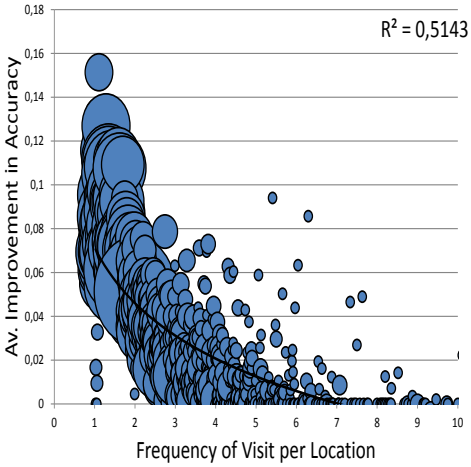

Figure 5.25: The plot shows a negative correlation between the frequency of visit per location and average absolute improvement in accuracy $r = -0.34, \rho = -0.80, P(\epsilon) = 0.0$.

5.5.4.4 User Entropy

The users in the Foursquare dataset were found to be very entropic, i.e. they visited many locations with similar probabilities. The average entropy of the users in the Foursquare dataset is 3.48 ± 0.71, which means, on average, a user can be expected to be in one of $2^{3.48} = 11.16$ locations. Thus the uncertainty of predicting the next location of a user in this dataset is very high. As stated in previous chapters, entropy is a very good indicator of the predictability of the mobility of a user. A low

entropic user, for example, has a few significant locations, which they regularly visit with different high probabilities, and possibly a few non-significant locations with low probabilities. A highly entropic user, in contrast, visits many locations with similar a probability, therefore prediction of their next location is subject to a high degree of uncertainty. The inclusion of the location histories of friends helps find more evidence (being with a friend) which can reduce the uncertainty during prediction of the next location. Indeed, the improvement in accuracy by incorporating the social network of the users shows a strong positive trend with the entropy of the users (figure (5.26)). The positive trend is confirmed by both Pearson's and Spearman's correlation coefficient values of $r = 0.38$ and $\rho = 0.45, P(\epsilon) = 0.0$ respectively. Figure (5.27) shows the relationship between entropy (x-axis) and average improvement in accuracy. The positive trend becomes more clear, the correlation between both quantities increase significantly. Pearson's correlation coefficient shows a very strong positive trend of $r = 0.64$. The positive trend is even stronger according to Spearman's rank correlation coefficient $\rho = 0.72$.

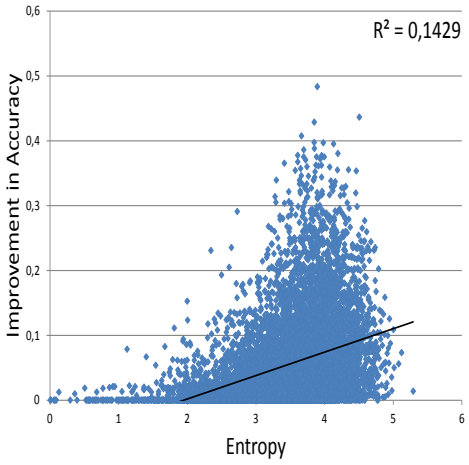

Figure 5.26: Absolute improvement in accuracy shows a positive trend with the increasing entropy of the users. Both Pearson's and Spearman's correlation coefficients are found to be $r = 0.38, \rho = 0.45, P(\epsilon) = 0.0$ respectively.

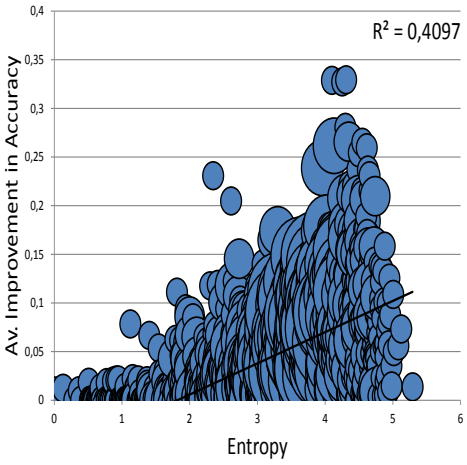

Figure 5.27: Average absolute improvement in accuracy shows a positive trend with increasing entropy of the users. Both Pearson's and Spearman's correlation coefficients are found to be $r = 0.64, \rho = 0.72, P(\epsilon) = 0.0$ respectively.

5.5.4.5 Location Entropy

Location entropy is a measure that quantifies the predictability of a location. A highly entropic location is visited by many users with a comparably similar frequency. Examples of highly entropic locations are airports, sport stadiums, underground stations, etc. A mediocre entropic location is visited by many users with different frequencies, for example, a restaurant in the neighborhood is visited frequently by the neighboring residents and sporadically by visitors from elsewhere. A low entropic location, for example, could be the private domicile of a user where friends come by occasionally. Therefore, location entropy is a good indicator of the pre-

dictability of a location. The lower the location entropy, the higher the predictability. The incorporation of location histories of friends increases the predictability of highly entropic locations, which are public locations accessible for everybody (see figure (4.1)). Figure 5.28 sets both improvement in accuracy (y-axis) and location entropy (x-axis) in relation to one another. Indeed, the improvement in accuracy increases as the location entropy increases. This positive trend is confirmed by a moderate positive correlation according to both Pearson's correlation coefficient $r = 0.23$ and Spearman's rank correlation coefficient $\rho = 0.23, P(\epsilon) = 0.0$. The relationship between average improvement in accuracy and location entropy on figure (5.29) show a very strong positive trend. Both Pearson's and Spearman's correlation coefficients increase significantly to $r = 0.57$ and $\rho = 0.60$, confirming the strong positive trend.

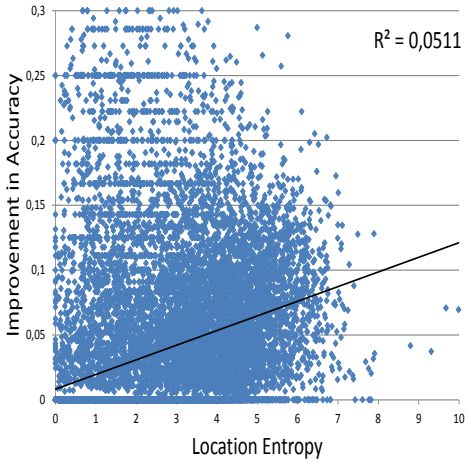

Figure 5.28: Absolute improvement in accuracy shows a positive tendency with the increasing entropy of the locations. Pearson's and Spearman's correlation coefficients were found to be $r = 0.23, \rho = 0.23, P(\epsilon) = 0.0$ respectively.

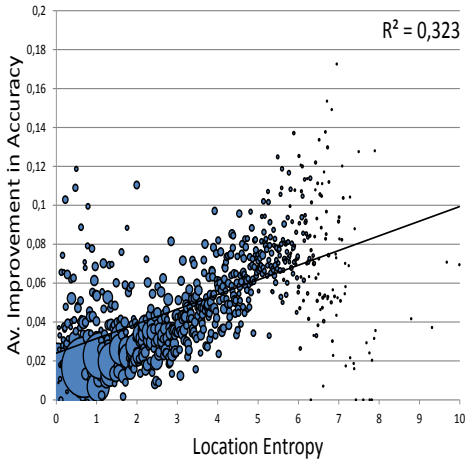

Figure 5.29: Average absolute improvement in accuracy shows a positive tendency with the increasing entropy of the users. Pearson's and Spearman's correlation coefficients were found to be $r = 0.57, \rho = 0.60, P(\epsilon) = 0.0$ respectively.

5.6 Mobility Models Based on Discrete HMMs

The Hidden Markov Model (HMM) is probably one of most popular, well-studied and powerful DBN approach (2.2.3.3). HMM is 5-tuple $\mathfrak{H} = (S; T; E; B; \pi)$, where S is the state space and represents the values of the random state variable x_t, T is a transition matrix (model) whose elements a_{ij} specify the probability of transiting from one state x_i to another x_j, E is the alphabet of possible observations, B is the emission matrix (model) matrix whose elements specify the probability of making an observation in a certain state $P(e_t|x_t)$, and finally π represents the initial probability distribution whose elements specify the probability that $\pi_i = P(x_1 = s_i)$ is the start state of the model at time step $t = 1$. HMM assumes that an observation $e_t \in E$ is generated by one component of a mixed distribution containing n components. Although the components are hidden, their number must be known a priori. The

components are modeled as a single latent state variable X of the underlying HMM, each component corresponds to a state of X.

Each of the observations $e_t \in E$ is related to a state $x_t \in X$ of the model at time step t. The conditional probability $p(e_t|x_t)$ forms the emission matrix B and represents the probability of an observation e_t being generated when the model is in state x_t. The transition matrix T holds the probability distribution of moving from one state to another $p(x_t|x_{t-1})$, thus its size is bounded by the state space $S \times S$. The states of the model are generally interconnected in such a way that any state can be reached from any other state [Rabiner, 1989]. Thus the number of states in the model and the number of observations are both critical issues of HMM, because the observations must be large enough to allow a transition between two arbitrary states. Otherwise the transition matrix contains zero entries meaning that some states cannot be reached from some other states and the model is said to suffer from the zero-frequency problem.

Given the initial state, state transition and both transition and emission model of an HMM, the unknown model parameters, i.e. the state transition matrix T and the emission matrix B can be learned using different algorithms, for example, the Baum-Welch learning algorithm [Baum et al., 1970] which uses the forward-backward algorithm [Russell and Norvig, 2010, P 579], [Duda et al., 2001, P 137-138]. Learning the exact parameters of an HMM and the complexity of the solution represents an intractable task, because as none of the known learning algorithms can find an exact solution, they tend to find a solution with local maxima [Ron and Singer, 1996]. Therefore, selection of the initial parameters of an HMM is of immense importance.

As stated earlier, the state space of a HMM must be known a priori, otherwise the proper state space of the model must either be guessed or learned. The state space can be guessed by testing different state spaces and choosing the state space that maximizes the accuracy of the model. Alternatively, the state space can be learned in a preprocessing step using different classifiers such as k-means (2.4.1.1) and DBScan (2.4.1.3) in order to find regions of high observation occurrence. Thus, the state space of an HMM is even more critical.

The initial state distribution π and both transition T and emission B models can be chosen randomly (for example, a uniform distribution), but learning using randomly chosen initial parameters is very time consuming. Further, there is no guarantee that the model will converge to a globally optimal solution instead of local maxima. To reduce the convergence time and to increase the probability of convergence to a true solution, it is very important to select good initial model parameters (for example, an a priori distribution can be estimated using the distribution of the observations over the states in order to determine the initial model parameters). Thus modeling an HMM requires in-depth understanding of the application domain.

An HMM can also be used for mobility modeling, the components of the model can be thought as states of mind or moods of a mobile user and observations as the locations they visit. An HMM such as this can be interpreted as follows, the user visits different locations according to their mood, thus the locations visited represent the outputs of their moods. The mood of a user and the number of different moods cannot be directly observed and are both unknown. An unknown number of

unknown states influences the movement behavior of a user, hence the term "hidden" in "Hidden Markov Model".

In [Jesdabodi, 2012], two different HMMs have been constructed, one model with and the other without considering influences from social networks. The first model uses the hours of the day and the day of the week as a state vector. The second model additionally uses a binary variable to indicate whether the user is accompanied by friends. Choosing a different number of states, the initial model parameters are estimated by clustering the sequence of observations using the k-means (2.4.1.1) clustering algorithm. Both models have been evaluated using the same dataset used for the evaluation of SOST PPM VOMM Model. In order to reduce the effects arising from the zero-frequency problem, evaluation is performed only for users who have at least 200 check-ins with a minimum of one check-in per day. The model's parameters are learned using the Baum-Welch algorithm. Table 5.3 reports the evaluation results using different numbers of states.

# States	Av. Accuracy	Av. Accuracy Social
2	16.4	16.6
5	15	15.2
10	14	14.1
20	12	12.1

Table 5.3: HMM accuracy in percent

The results show that the integration of the social network leads to a non-significant absolute improvement in accuracy ranging from 0.001 to 0.002. This improvement in accuracy is significantly less than the improvement in accuracy achieved when incorporating social influences in an SOST PPM VOMM. Further, the accuracies of both HMM models are significantly less than the accuracy of the spatial-temporal PPM VOMM model. The poorer results of the two HMM models are not due to an HMM being less powerful, but are rather due to the difficulty of modeling and learning the model parameters.

Further, the results show that accuracy decreases as the number of states increases for one of two reasons, either the data has been drawn from one distribution (the users have only one state of mind), or the location histories of the users are insufficient to learn the exact model parameters. The second reason seems more probable, as observation size is known to be a critical issue for HMM [Ron and Singer, 1996, Begleiter et al., 2004]. The assumption that a larger history size contributes to improvement in accuracy is partially correct, because there is no history size which guarantees that a transition between two arbitrary states will occur, and even if this were so, the larger history size could lead to other problems such as overfitting (the model describes artifacts of the dataset rather than the true solution). A larger location history may lead to a changed mix of distributions (figure 5.20 and 5.18 show that the entropy and the number of locations visited by a user increases as the history size increases), this in turn requires the model to be re-trained, which is another critical issue of HMM. Even if the number of states does not change, a larger history size leads to a severe readjustment of the model parameters, especially when frequently occurring states are no longer relevant for a user (for example, behavioral

changes due to marriage, moving or a new job, etc.).

Finally, HMM uses one state variable, thus, incorporation of new information sources is possible by increasing the dimensionality of the state variable. The state space of an HMM model grows multiplicatively with the dimensionality of the state space, which implies the need for even more training data.

Chapter 6

Additional Context (AC) & Discrete Knowledge (DK)

Contextual data is not limited to spatial, temporal or social features. Nowadays mobile devices are equipped with a set of additional sensors that can deliver valuable information about the behavior of a user. Examples of such sensors are accelerometers, W-LAN, Bluetooth, system programs, GSM sensors, etc. Beside these sensors, additional sources of information exist that may provide information that can help detect temporary deviations from the usual routines of a user. For example, the calendar function may contain information about appointments, indicating that a user most probably is abandoning their normal routine temporarily, but they will return to their routine behavior after a certain time period elapses. We classify this type of information as discrete knowledge, because the information is only temporarily valid. It has a short life time beyond which influence on the behavior of the user is lost (unlike contextual knowledge which is valid for a longer period of time).

Chapter summary: Section I provides an introduction to additional information sources, the integration of additional information sources into the mobility model, related work, the MDC dataset and empirical results. Discrete knowledge, people's adherence to schedule, the extension of the mobility model in order to incorporate discrete knowledge and finally empirical results are presented in section II.

6.1 Additional context

The rapid development of both hardware (smart phones equipped with numerous sensors) and software (mobile applications - APPs) has opened up new perspectives as well as challenges in research into human behavior generally and human mobility specifically. The new integrated smart phone features and mobile applications allow huge amounts of data to be collected, this is data such as W-LAN, GPS and GSM observations, as well as call logs, SMS, Bluetooth use, clock alarms, user profiles, calendar entries, etc.

Different types of data provide different information about different behavioral aspects of a mobile user. For example, WLAM, GSM and GPS provide data about the position of the mobile user. Bluetooth, call logs and SMS provide information about exchanges of information or communication among the users, whereas a user's personal calendar, clock alarm or device settings provide information about scheduled tasks of the user at a certain point of time. All these data types can either be directly related to a location, i.e. W-LAN, GSM and GPS, or be indirect evidence for being at a certain location, for example the clock alarm setting is a indirect indication of the user being at a place where they can relax/sleep and setting the system profile of the device to silent is an indication they may be entering a meeting at a work place or a doctor's surgery, for example. Bluetooth communication may also indicate the user is with friends, which in turn may be associated with being at certain locations, such as the home of a friend or their favorite bar.

Formally, let $A = A_1, A_2, A_3, A_4, ..., A_i$ be the set of additional information sources. Assuming conditional independence between the different information sources, the probability of the user being at location $q \in \Sigma_{loc}$ given the current spatial context s, time t and additional observations $O_t = o_1 \in A_1, o_2 \in A_2, ..., A_j \in A_j$ from the sensors contained in A can be calculated according to equation (6.1).

$$\begin{array}{rcl} p(q|s,t,O_t) & = & \frac{p(s,q,t,O_t)}{p(s,t,O_t)} = \frac{p(O_t,s,q,t)}{p(O_t,s,t)} \\ & = & \frac{p(O_t|s,q,t)}{p(O_t|s,t)} = \prod_{o_i \in a_i} \frac{p(o_i|s,q,t)}{p(o_i|s,t)} \end{array} \tag{6.1}$$

where $q \in \Sigma_{loc}$ is a possible future location of the user. If we assume O_t depends on the current location only then the probability $p(q|s,t,O_t)$ changes according to equation (6.2).

$$\begin{array}{rcl} p(q|s,t,O_t) & = & \frac{p(q,O_t|s,t)}{p(O_t)} \\ p(q,O_t|s,t) & = & p(q|s,t) * p(O_t|s,q,t) \\ & = & p(q|s,t) * p(O_t|q) \\ & = & p(q|s,t) * \prod_{o_i \in a_i} p(o_i|q) \end{array} \tag{6.2}$$

The future location q of the user is the location that maximizes the probability $p(q|s,t,O_t)$:

$$q = \text{argmax}_q p(q|s,t,O_t) \tag{6.3}$$

If the set of observations at prediction time t is empty then the model switches back to the spatial-temporal model detailed in the previous chapters.

6.1.1 Additional Context (AC) PPM VOMM Tree

The extension of the PPM VOMM tree must enable estimate of probabilities in the form $p(o_i|s,q,t)$, where $o_i \in A_i$ is an observation from the information source $A_i \in A$, q is a candidate future location, s is the current spatial context and t is the current time step. Each node in the AC PPM VOMM tree corresponds to a spatial-temporal context q,s,t, thus we just need to extend all nodes so that they hold counters for each observation $o_i \in A_i$ that count the occurrence of patterns of the form o_i,s,q,t. The counters for both q,s,t and o_i,s,q,t allow estimate of the probability $p(o_i|s,q,t)$ according to equation (6.4):

$$p(o_i|s,q,t) = \frac{C_{o_i}^{(A_i)}(s,q,t)}{C(s,q,t)} \tag{6.4}$$

Where $C_{o_i}^{(A_i)}$ is an additional counter for counting the occurrence of observation o_i from information source A_i and $C(s,q,t)$ is the spatial-temporal counter as in the previous chapters. AC PPM VOMM manages the additional counters in the nodes of both spatial tree and temporal sub-tree. The probability $p(o_i|q)$ can be estimated from the additional counters as follows:

$$p(o_i|q) = C_{o_i}^{(A_i)}(q)\frac{C_{o_i}^{(A_i)}(\epsilon)}{\Sigma_{a_k \in A_i} C_{a_k}^{(A_i)}(\epsilon)} \tag{6.5}$$

where ϵ represents the root of the VOMM tree and $C_{a_k}^{(A_i)}(\epsilon)$ is the counter that holds the occurrence of observation $a_k \in A_i$ at the root node ϵ.

Extending the PPM VOMM tree to integrate additional information sources is effectively adding counters for the observations of each new information source in order to hold the occurrence of the observations according to the spatial-temporal context represented by the node itself.

6.1.2 Related Work

The winner of Nokia Data Challenge (MDC) [Etter et al., 2012] has compared different mobility model approaches and combined them using a blending strategy aimed at improving the prediction accuracy. The authors report an accuracy of $60.07\%, 60.83\%$ and 57.63% for three mobility models which are based on a Dynamical Bayesian Network (DBN), Artificial Neural Networks (ANN) and Gradient Boosted Decision Trees (GBDT) respectively. The authors used only spatial and temporal features and reported no improvement when incorporating additional contextual knowledge such as GSM or W-LAN.

The second winner of MDC was [Wang and Prabhala, 2012]. The authors predict the next location of a user by combining a 1-order Markov Model and a Periodicity Based Model PDM. The next location of the user is predicted by the 1-order Markov Model as long as the probability exceeds a threshold, otherwise the PDM model is used. The PDM model determines dense clusters of visits in time for each location,

i.e. the most frequent periods of the week when the user visits the location. The next location is the location with the largest cluster according to the current (prediction) time. The authors used only spatial and temporal features and report their best prediction accuracy to be 55.69%.

The third winner of MDC was [Gao et al., 2012]. They used a mobility model called Hierarchical Pitman-Yor (HPY), which was originally used in language processing. The HPY mobility model considers both spatial and temporal information for predicting the future location of the user. Assuming that the time of a visit depends only on the current location (not the whole spatial context) and assuming that the time of visits to a location is a Gaussian distribution, they calculate the probability of a visit to a location l and time t according to the following equation:

$$p(l_i|t, l_{i-1}) = p(l_i|l_{i-1}) * \prod_{\tau \in \lambda} \mathcal{N}_{l_i}(\tau|\mu_\tau, \sigma_\tau^2) \tag{6.6}$$

where l_i is the location of the $i-th$ visit, λ contains the set of temporal features extracted from time t (such as the hour of day and the day of week) and $\mathcal{N}_l(\tau|\mu_\tau, \sigma_\tau^2)$ represents the Gaussian distribution of the visits to location l_i over the temporal feature $\tau \in \lambda$. The authors report that the mobility model that considers both temporal features "hour of day" and "day of week" has the best prediction accuracy of 50.53%. Assuming the distribution of visits over the hours of day and the days of week to be Gaussian is less applicable in this case, for example, consider a student who visits a certain course at the same lecture hall every Monday afternoon and every Thursday morning, it is clear that the visits of the student to the lecture hall are a mixture of two distributions.

An approach based on user-specific decision trees learned from each user's history has been proposed in [Tran et al., 2012] in order to predict the next location of the user. The authors report an accuracy of 61.11% after preprocessing the dataset for detecting holidays, Home and Working locations and huge gaps between consecutive data points in order to interrupt the current trajectory sequence. A weakness of this model is the elaborate pre-processing, which is most possibly responsible for the good results.

6.1.3 Evaluation

The mobility models presented in the related work subsection were evaluated using the MDC dataset [Laurila et al., 2012]. We evaluate our AC PPM VOMM mobility model based on the same dataset, which allows us to compare the performance of our mobility model with the performance of these mobility models.

6.1.3.1 Nokia Data Challenge (MDC) Dataset

"The Mobile Data Challenge (MDC), a large-scale research initiative aimed at generating innovations around smart phone-based research, as well as community-based evaluation of related mobile data analysis methodologies." [Laurila et al., 2012]. The Lausanne Data Collection Campaign (LDCC) is an initiative started to collect

a unique, longitudinal smart phone dataset for the basis of MDC [Laurila et al., 2012]. The LDCC dataset contains quasi-continuous measurements covering all sensory and other available information from a smart phone [Laurila et al., 2012]. Table (6.1) contains statistics of LDCC relevant for AC PPM VOMM model that incorporates additional contextual knowledge.

Data type	**Quantity**
# Users:	80
# Location points:	26 152 673
# Unique cell towers (GSM):	99 166
# Application events:	8 096 870
# Bluetooth observations:	38 259 550
# Unique Bluetooth devices:	498 593
# W-LAN observations:	31 013 270
# Unique W-LAN access points:	560 441
Av. unique locations per user:	66.30 ± 33.90
Av. history size:	573.66 ± 324.33
Av. history size per location:	8.78 ± 3.92
Av. entropy:	2.42 ± 0.39

Table 6.1: An extract from the MDC dataset statistics provided in part by [Laurila et al., 2012] or calculated during our further analysis of the dataset.

The dataset is strongly anonymized so that the construction of any joint models for improving the prediction accuracy for one user based on knowledge from other users is impossible. The dataset is also rather sparse (the average history size per user for 14 months is just 573.66 ± 324.33) and contains huge gaps with no observations, for example some users only have observations during the start and the end of the data collection period, but no observations in the intervening time.

6.1.3.2 Feature Extraction

Additional information sources can contain features related to different aspects of the contexts of the user. Feature selection is an important task in constructing predictive models. The aim of feature selection is to remove irrelevant and redundant features and to select only those features that are relevant for the model (taking into account how these features are related). The selection of relevant features influences the performance and interpretability of the model, it reduces the training time and avoids overfitting. We use features relating to the spatial, social and individual contexts of the user. The following list shows the information sources we used as additional information sources:

- **GSM** G **-** contains the anonymized GSM cell IDs that the user has seen. This additional feature provides further (coarser) information about the location of the mobile users.
- **W-LAN** WL **-** contains the anonymized MAC-addresses of the W-LAN devices seen by the device of a user. W-LAN observations also relate to the

spatial context of a user. Its values are evidence for the location of W-LAN access points.

- **System settings** S **-** contains the system profile selected by a user(for example silent, general, etc.) and the type of ring as "normal", "ascending", "ring once", "beep" or "silent". Both profile and ring settings are related to the individual context of a user, for example a user sets their ring to "beep" whenever they go to bed or a user sets their system profile to silent whenever they enter certain locations (a meeting room, doctor's surgery). Thus both profile and ring settings may be evidence for being at such locations.
- **Bluetooth** B **-** contains the anonymized MAC-addresses of the Bluetooth devices seen by the device of a user. The observations from the Bluetooth device relate to the social context of the user. Bluetooth covers a small area ranging between 1 and 100 meters, therefore the observation of other devices means that the users have met.

Table (6.2) provides the set of features to be included in the AC PPM VOMM model:

Variables	**Domain**	**Description**
Σ_{loc}	$\{l_1, l_2, ..., l_i\}$	The set of locations visited by the user
W	$\{Wd, We\}$	a binary variable representing whether it is a weekend day or a work day
D	$\{Sun, Mon, ..., Sat\}$	The day of week
$S^{\Delta t}$	$\{S_1, S_2, ..., S_j\}$	The number of time slots calculated by dividing the hours of day by Δt, setting $\Delta t = 1$ means the each hour of day represents a slot
W	$\{0, 1\}$	a binary variable representing whether it is a week end day or a working day
G	$\{g_1, g_2, ..., g_s\}$	The set of cell tower IDs
B	$\{b_1, b_2, ..., b_k\}$	MAC-addresses of Bluetooth devices observed by the mobile device of the user
WL	$\{w_1, w_2, ..., w_l\}$	MAC-addresses of W-LAN devices observed by the mobile device of the user
S	$\{s_1, s_2, ..., s_m\}$	The set of device settings of the mobile device of the user, each setting is a combination of current ring and active profile

Table 6.2: The list of variables and their domains that are used by AC PPM VPMM.

According to the MDC instructions, a time interval is defined as follows. The start point of an interval is set to ten minutes before the start of a visit and the end point of the interval is set to the end of that visit [Laurila et al., 2012]. The competitors are allowed to use whatsoever data they wish within an interval to build a context. We use the observations made within the last 30 Minutes of an interval for building the context because we believe that the last 30 minutes of observations have the most impact on the next visit of the user. A nice side effect of our decision is the reduction in the amount of data to be considered at any time step.

6.1.3.3 Empirical Results

The maximum accuracy of the AC PPM VOMM is 60.63%. The performance of AC PPM VOMM is as good as the mobility models introduced in the related work subsection. A comparison between the models is reliable, because all models are evaluated using the same dataset. We have applied our model to the raw data and omitted any pre-processing (for example, for the detection of holidays and user-specific off-days). Moreover, we have chosen the parameters of our model to be global and not user-specific (for example, user-specific drift functions or user-specific subdivision of the hours of the day into time slots). Therefore, we believe that the AC PPM VOMM has the potential to achieve an even better performance and that it can outperform the other models.

The inclusion of Bluetooth, W-LAN and system information results in an improvement in accuracy for some users, but the overall improvement in accuracy is negligible. Hence, we state that their inclusion does not contribute to an improvement in the performance of the model. Both system information and W-LAN are tightly related to the actual location of the user and do not provide any evidence about the future location of the user. Therefore these information sources may contribute to the prediction of the current location of the user, but not to the prediction of the next location. A user sets their system profile to silent when they enter a doctor's surgery or a meeting room and not much earlier, for example they do not change their system profile in advance at their previous location. A W-LAN access point can only be observed by the mobile device of the user when they arrive at the location and not earlier. Furthermore, Bluetooth observations are dependent on the presence of other devices in the neighborhood, which means the presence of other users at the current location of the user. Therefore the contribution of Bluetooth is highly dependent on knowledge about the social network of the users. Unfortunately we did not have access to the social network, therefore using Bluetooth contributes no significant improvement in accuracy.

Features	ST PPM VOMM	AC PPM VOMM DEP	AC PPM VOMM INDEP
Accuracy %:	58.11	58.54	60.63
Abs. Improvement:	-	0.0043	0.0252
Rel. Improvement %:	-	0.74%	4.34%
Two-Sided Unpaired T-Test $p(\epsilon)$:	-	0.3442	0.0638

Table 6.3: Empirical results: Column 2 represents the basic spatial-temporal ST PPM VOMM model, column 3 the model that includes additional information sources assuming dependency on the whole spatial context s, column 4 the model that includes additional information sources assuming dependency on the most recent location.

Table (6.3) shows the experimental results of evaluating AC PPM VOMM taking only GSM observations into account. The inclusion of GSM leads to an absolute improvement in accuracy of 0.0043 and a relative improvement in accuracy of 0.74% assuming dependency between GSM and the whole spatial context s and an absolute improvement in accuracy of 0.0252 and a relative improvement in accuracy of 4.34% assuming that GSM depends only on the current location rather than the whole

spatial context s. The higher improvement in accuracy for the later model is an indicator that GSM does indeed only depend on the current location of the user.

6.2 Discrete Knowledge (DK)

Discrete knowledge is a unit of evidence about the occurrence of an event at a certain point of time. The evidence expires after the point in time elapses.Discrete knowledge represents temporal interruptions of standard movement patterns (daily routine) of a user for a certain period of time before returning back to the standard routines. In contrast to the features mentioned previously such as time, space, social, etc. discrete knowledge has no continuous influence on the behavior of a user. Discrete knowledge has no history of observations that could help to infer the mobility behavior of the user beyond the expiration date. Examples of discrete knowledge are appointments in the personal calendar of a user, cinema or restaurant reservations made over a web browser or via a mobile phone, news feeds from the local police or city authorities about events, blocked routes or metro timetables, etc.

As already stated, discrete knowledge has no observation history and no context, thus no patterns can be detected. Discrete knowledge helps detect a preliminary interruption of a standard routine of a user, when the user makes an unusual movement at a certain point in time, but then returns to their usual routine after this time has lapsed. For example, when a user visits a doctor after work instead of going straight home as usual. Such an unusual movement does not cause only a false prediction of the next location, but it also causes n false future predictions (according to the order of the underlying mobility model). Therefore integration of discrete knowledge during the prediction task can improve the accuracy of the mobility model. Furthermore, discrete knowledge can contribute to the prediction of locations where a user has never been before, for example such as the location of an event or an appointment with a new doctor.

In the following subsection we assume that discrete knowledge is scheduled, i.e. the start point in time is known, and that it is always be associated with known locations. Discrete knowledge that is not scheduled or cannot be associated with a location is beyond the scope of this chapter.

6.2.1 Adherence to Schedules

The influence of discrete knowledge on the mobility of a user depends on the user's adherence to appointments. Adherence to schedule varies from user to user. Whereas some users exhibit a very high adherence to schedules of any kind, the behavior of others depends on the type of the schedule. Users are usually vigilant about keeping medical appointments, but some users take cinema or restaurant reservations less seriously. It is tempting to assume a user would not bother to make an appointment if they did not intend to keep the appointment, nevertheless some users have a lax attitude to keeping their appointments.

Due to the lack of a dataset that allows an individual's adherence to schedules to be estimated, we used another method, namely a survey. A survey allows a global

estimation of people's adherence to schedules. We called a number of cinemas, restaurants, medical offices, beauty and hairdressing saloons and simply asked for data on how often customers kept their appointments. In addition, we asked for the percentage of users who cancelled their appointments if they could not keep them. We categorized the schedules into different types. We distinguished between food, entertainment, health care and others. Table (6.4) shows the average results of the survey.

Schedule type	Compliance rate	Cancelling rate
Food	0.92	0.60
Health Care	0.95	0.90
Entertainment	0.90	0.60
Others	0.85	0.60

Table 6.4: The results of the survey on users' adherence to schedules.

We used the results of the survey as the initial values for the adherence to schedules for each user in following subsections.

6.2.2 Integrating Discrete Knowledge (DK) into PPM VOMM

We assume a user's adherence to schedules is known at any time step. Further, we assume the user's time schedule is known at any time step. Upon notification of the existence of a schedule, we temporarily boost the probability of the location associated with the schedule by multiplying the probability of this location being the next location of the user $p(q|s,t)$ with the user's adherence to schedules ϑ. The probability of the remaining locations is then dampened by the factor $1-\vartheta$.

For each type of schedule we have two counters, one for saving the occurrence of schedules of the associated type and the second counter for schedules, i.e. appointments that have been kept by the user. Initially we set the first counter to 100 and the second counter is calculated by multiplying 100 with initial value from the survey. We update the counters according to futures occurrences of schedules, thus the initial values of the respective types are adapted according to the user's adherence to schedules.

6.2.3 Empirical Results

Due to the lack of a dataset enriched with schedules, we made use of the Reality Mining dataset. The users label the cell tower IDs when they have visited a significant location. Although the labels are user specific, the users often use known names and addresses as labels. Thus, some labels can be associated with real locations such as restaurants, doctors, theaters, cinemas, bars, hospitals, etc. We enriched the Reality Mining dataset with artificially created schedules for around 2% of the visits of all users. We selected those visits that can obviously be associated with locations where people usually arrange an appointment or make a reservation prior to their visit.

The accuracy of an order 3 ST PPM VOMM is 82.65% without consideration being given to discrete knowledge. DK PPM VOMM is the mobility model which considers discrete knowledge. Prediction accuracy increases to 83.81% when applying DK PPM VOMM to the Reality Mining dataset enriched with schedules. The absolute improvement in accuracy compared to ST PPM VOMM corresponds to 0.012 and the relative improvement in accuracy corresponds to 1.4%. The significance of the improvement is confirmed by a two-sided unpaired Student's t-test with a value of 0.038.

Chapter 7

Conclusion

Location prediction has been the subject of many investigations and various approaches have been applied to building mobility models, such as random, computational and probabilistic reasoning. Random mobility models are not able to model realistic human movement, because human movement is not random and follows a mobility pattern. However, mobility models which can calculate the exact location of the user based on physical laws such as distance, velocity and acceleration are suitable for short-term predictions of a few seconds. Long-term predictions of a few hours to a few days into the future are subject to high uncertainty, because a user can travel long distances spanning hundreds of kilometers within a time period of only a few hours or a few days, thus the computational estimation of the future location of a user is almost impossible. However, mobility models based on probabilistic reasoning can handle such uncertainty and appear to be well-suited for long-term predictions.

Mobility models based on probabilistic reasoning predict the future location of users given their location histories. Probabilistic estimation of the future location of a user is unfeasible using the complete joint probability over the whole location history, therefore, a Markov assumption is often used to simplify the probability estimation of the user being at one particular location. The Markov assumption states that at any time step, the influence of the whole history on the probability of the user being at a location is included in the n most recent observations (visits). Thus, the n most recent observations are sufficient for estimating this probability, the underlying mobility model is then said to be of the order n.

Mobility models based on probabilistic reasoning usually consist of two steps, namely the training and the prediction step. The training step uses the location history of the user to detect as many movement patterns as possible. The length of the patterns depends on the order of the mobility model n. The prediction phase is used for assigning a probability mass to each location in the location history of the user given the movement patterns and the current n observations of the user. The model predicts the location with the highest probability mass.

The patterns detected in the location history of a user vary according to their length and dimensionality. The length of a pattern is bounded by the order of the underlying mobility model, whereas dimensionality is defined by the features

included in the mobility model. A feature can be related to the spatial, temporal, social or additional context of the user. The number of features and the size of the pattern detected have a direct influence on both the training and prediction steps of the underlying mobility model. The critical issues related to mobility models based on probabilistic reasoning were addressed in previous chapters, most critical issues are both cold-start and zero-frequency problems, as well as the size of location history, the frequency of visit per location and the entropy of the users. The focus of this work has been on looking at methods for alleviating the drawbacks of these models and improving their performance by integrating features related to location, time, social network, additional context and discrete knowledge into the mobility model.

7.1 Mobility Model & Empirical Results

The following subsections provide a summary of the results and findings of each chapter.

7.1.1 The Mobility Model Approach

Two promising approaches based on probabilistic reasoning are the Fixed Order Markov Model (FOMM) and the Hidden Markov Model (HMM). FOMM treats the whole location history as a single distribution, therefore, FOMM consists of one (observable) parameter, namely the transition matrix for estimating the probability of moving on to a location given a context s of length $n = |s|$. The size of the transition matrix grows exponentially with the order of the model. In contrast to FOMM, HMM treats the location history as a mixed distribution of c unknown components. Thus HMM consist of two parts, namely the hidden (or latent) state model, which controls the component from which an observation (output) is drawn, and an observable emission model, which represents the output (for example, the future locations of the user) of the model. The Markov assumption is applied to the state model in HMM, which is managed in a transition matrix. The emission probabilities are managed in the emission matrix. A pattern in HMM consists of a sequence of n states. HMM is much more flexible than FOMM, but training HMM is a non-trivial task, which has not yet been solved by any of the learning techniques.

Both FOMM and HMM have a fixed order, thus the patterns detected by them have a fixed size. Choosing a higher order n for the model increases the number of patterns that could be detected, but at the same time increases the need for training data. Thus both models suffer from the cold-start and the zero-frequency problem. A predictive model suffers from cold-start when insufficient data is available for training the model. A model suffers from zero-frequency problem when the training data does not contain all the patterns that can be detected by the model, thus the a priori probability of a pattern that does not occur in the location history is zero. An a priori probability of zero causes the posteriori (prediction) probability to be zero. A posterior probability of zero based on the a priori probability is far from reality, because a user can visit a location for the first time at any point in time in the future. The zero-frequency problem can be alleviated by using more training data,

but there is no guarantee that more training data will contain all possible patterns, thus the zero-frequency problem remains unresolved.

Human behavior is continuously subject to change over time, therefore a mobility model has to be able to adapt to the changing behavior of the user. The adaptability to the changing behavior of users is another problem of FOMM and especially HMM. When the frequent patterns of a user change, the model has to be retrained, which is costly and requires a lot of effort. Another problem of HMM is its training, which is non-trivial as stated earlier. Selection of the initial parameters is very important so that HMM learning converges to a true solution. Therefore, training HMM requires in-depth understanding of the application domain so that suitable initial parameters can be selected. Furthermore, it is questionable as to how much extra performance is gained when the initial parameters are near to the true solution. In many cases, the initial parameters can be sufficient for achieving reasonable performance and consequently the hidden parts of HMM just complicate the model unnecessarily, hence the arduous journey of training an HMM can be spared.

7.1.2 Spatial Context

We use an approach based on context specific Bayesian networks, more specifically we use a variant of the Variable Order Markov Model (VOMM) called Prediction by Partial Matching (PPM). The PPM VOMM mobility model has a variable order, the model switches automatically to a lower order if it cannot find a pattern of length equal to the higher order. PPM VOMM uses an escape mechanism for switching to a lower order. The escape mechanism of PPM VOMM is a Laplace-like estimator for assigning a probability mass to patterns that do not appear in the location history. Thus PPM VOMM is able to alleviate the drawbacks of zero-frequency to a high extent. The variable order of PPM VOMM allows detection of patterns of variable size between one and n, therefore PPM VOMM is able to detect significantly more patterns in the same amount of location history. Hence, PPM VOMM is less dependent on the size of the location history and suffers less from cold-start or zero-frequency problems.

PPM VOMM is structurally much simpler than HMM, it manages the patterns in a tree structure. Each node of the tree is labeled with a symbol from the (spatial) alphabet and corresponds to a pattern constructed by traversing the tree from the root to that specific node and concatenating the label of all nodes on that path. Each node has a counter for bookkeeping its occurrence. Learning PPM VOMM is actually the process of inserting paths into the tree or updating the node counters. Therefore PPM VOMM allows both the training and prediction phases to be conducted simultaneously, which further alleviates both cold-start and zero-frequency problems and reduces the dependency of PPM VOMM on the quantity of location history.

We have tested the performance of PPM VOMM using different datasets, namely Reality Mining [Eagle and Pentland, 2006], GeoLife [Geo, 2013a, Geo, 2013b], Foursquare [Fou, 2012b] and Mobile Data Challenge (MDC) [Laurila et al., 2012] dataset. PPM VOMM outperformed FOMM for all datasets. The datasets we used have also been used by other researchers to evaluate other mobility model approaches. The mobil-

ity model approaches were based on Dynamic Bayesian Networks, Artificial Neural Networks, Principal Component Analysis, Gradient Boosted Decision Trees, Hierarchical Pitman-Yor (HPY), etc. PPM VOMM outperforms, or at least is as good as the accuracies reported in other works.

Extensive empirical analysis demonstrates the significance of the improvements in accuracy using PPM VOMM compared to FOMM (up to 18%). The mobility model based on PPM VOMM is less dependent on the size of training data or the order of the model. PPM VOMM has significant advantages when the user travels in unknown terrain, where he has never been before. Although PPM VOMM is not yet able to predict locations visited for the first time, it is able to predict the future location of a user even if the n most recent locations of the user (current spatial context) contain unknown locations. PPM VOMM was able to predict the next location of the user in almost 48.84% of the cases when the current context of the user contains unknown locations. This corresponds to an overall improvement in accuracy of 11.59%. The distribution of the improvements over the hours of work and weekend days shows that most of the improvements occur during the morning hours of work days prior to arriving at the work location, evening hours after leaving the work location and during the evening and night hours of weekend days. People are most explorative and visit new locations during the aforementioned hours, which underlines the advantages of both the variable order and escape mechanism of PPM VOMM for predicting the next location of a user during the most interesting (for service providers) hours of the week.

The improvement in accuracy gained by using PPM VOMM compared to FOMM correlates negatively with the frequency of visits per location and positively with the number of locations visited by the user, hence PPM VOMM is less dependent on either the number of locations or the frequency of visits per location compared to FOMM. The performance of both PPM VOMM and FOMM becomes similar for locations visited with a very high frequency. Furthermore, PPM VOMM shows a better performance for highly entropic users who visit a lot of locations. The mobility of those users is usually associated with high uncertainty, because these users are very explorative in nature and consequently their location histories do not have dominant mobility patterns. The empirical results have impressively underlined the advantages of PPM VOMM compared to FOMM.

7.1.3 Temporal Context

The mobility of humans also obeys temporal regularities besides spatial-pattern. Humans visit many locations periodically, for example, many users go to their work locations every work day morning and they go home every evening between 6 and 8 p.m. The periodic behavior of humans is defined by hierarchically structured temporal features. Temporal features correspond to time units such as week of the year, day of the week, hour of day, etc. Temporal features follow an inclusion semantic, i.e. a temporal feature at level i in the hierarchy consists of n units of the (more specific) temporal feature in the hierarchy level $i + 1$. For example a day of the week consists of 24 hours in that day. Inclusion of temporal features in the mobility model helps detect more patterns. These patterns can be categorized according to

their dependency on spatial and/or temporal features into pure spatial-pattern, pure temporal-pattern and mixed spatial-temporal pattern.

The inclusion of more features in a mobility model increases the state space of the model. Including a new feature in HMM or FOMM causes both the transition and/or emission matrices to grow exponentially to the domain of the new feature. Thus inclusion of more features increases the need for more location history and the dependency of the mobility model on the size of training data. Both the variable order and the escape mechanism of PPM VOMM allow the inclusion of more features without increasing the need for more training data. We refer to the PPM VOMM model enriched with temporal features as the Spatial-Temporal (ST) PPM VOMM. The empirical results show that inclusion of temporal features increases the accuracy of the model using the same amount of training data up to 7% compared to PPM VOMM based only on spatial features and up to 25% compared to FOMM.

The empirical results show that the movement of a user to the next location and the amount of time they spend there depends on the arrival time at that location. Thus the mobility of a user is conditionally dependent on the temporal features extracted from previous locations of the user, thus it is conditionally independent from the spatial context s of the user. As stated earlier, the mobility behavior of a user is subject to continuous change over time. ST PPM VOMM uses a drift function in order to let non-recurrent patterns decay faster and to increase the adaptability of the model. The use of the drift function increases the performance of the model by almost 1.29%. The patterns of a user decay by factor of 0.007 every 6 hours, which means the patterns of a user have a life cycle of two years on average. On average, the influence of a mobility pattern on the movements of a user vanishes after two years of non-occurrence.

Extensive empirical analysis shows that inclusion of temporal features in an ST PPM VOMM significantly alleviates the drawbacks of cold-start and zero-frequency problems. When the current context of the user is unknown accuracy increases to 61.18%, which corresponds to a total accuracy improvement of 3.08% compared to PPM VOMM and almost 14.67% when compared to FOMM. ST PMM VOMM has further advantages over PPM VOMM in that it is able to detect pure temporal and mixed spatial-temporal patterns, thus it can further decrease the drawbacks of the zero-frequency problem. Again most improvements occur during the evening and night hours and during the morning hours of work days, when people are most active and explorative. As stated earlier, these hours are most attractive for service providers, because people usually conduct their free time activities during these hours.

An in-depth correlation analysis emphasizes the importance of temporal patterns for enhancing prediction accuracy. The correlation analysis shows that the ST PPM VOMM is less dependent on the size of training data and the frequency of visits per location. The improvements correlate negatively with both measurements. Further, the improvements in accuracy correlate even more strongly positively with both number of locations and entropy, which again confirms our finding that ST PPM VOMM is well-suited for predicting the mobility of users during their most active/explorative hours. ST PPM VOMM is able to find more patterns in the same amount of location history and thus is able to predict the even less frequent mobility

patterns of users such as locations where they spend their free time.

7.1.4 Mobile Homophily

The mobility of humans cannot be explained by focusing solely on the individual location history of a user. Humans are social beings and are subject to social influences. They influence and are influenced by others, who are socially connected to them. The mobility behavior of an individual is not excluded from social influences. Mobile homophily is the tendency of similar individuals to be interested in the same locations. Mobile homophily implies an interdependency between two quantities, namely social and mobile proximity. Social proximity measures the overlap between the social relationships of two users. We calculate social proximity based on measurements from social network analysis (SNA) like common neighbor, Jaccard Coefficient and Admic & Adair. Mobile proximity quantifies the similarity between the mobility of two users, i.e. to what extent two users are interested in the same locations. We calculate mobile proximity by quantifying the amount of spatial and spatial-temporal overlap between the location histories of two users. Additionally we have calculated a weighted version of both spatial and spatial-temporal overlap using measurements such as entropy, the distance travelled from the home location to a meeting location, the number of users who visit a location (in order to differentiate between public locations and private domiciles), and location density, i.e. the number of other locations in the neighborhood of the meeting location (because two users may be at the same location by chance if the users do not have a lot of location choice such as in a suburban area, whereas in urban areas people have a lot of choice, thus they visit the same location intentionally).

We used data collected from the online location -based social networking platform Foursquare between 23.03 and 23.07.2012, for demonstrating the interdependency between social and mobile proximities based on an extensive correlation analysis between network and mobility measurements. The dataset contained the location histories of 113,000 users, the social network contains approximately $7.5 * 10^6$ users. The social network has a clustering coefficient of 0.104 and an average shortest path) of 4.15, which confirmed that the users were building a real social network.

The correlation analysis shows a weak to moderate correlation between social and mobile proximities. The low correlation is due to the nature of online communities (Foursquare is one), social ties in online communities can be formed regardless of geographical barriers. The propinquity effect states that close things influence each other more than distant things. In line with this theory, we repeated the correlation analysis between users from the same city. The new correlation analysis showed a moderate to strong correlation between social and mobile proximities.

The majority of users interact mainly with a small group of their acquaintances. The acquaintances in the group exhibit a high degree of cohesion, which means that each member of the group spends time and interacts with almost all the other members of the group. The members of the same subgroups exhibit a high degree of similarity in their beliefs, thoughts, norms, goals, movements, emotional needs, etc. thus they are connected together will strong ties. We detected more than $8,700$ maximal cliques and more than $29,000$ maximal 2-plexes in the dataset. The av-

erage size of the maximal 2-plexes was 7.79 ± 2.41 and the largest 2-plex contains 20 users, which is in perfect agreement with the human social perception limit in [Fischer and Wiswede, 1997]. For each maximal 2-plex we calculated a measure of cohesion according to equation (4.2). A correlation analysis between group cohesion and mobile proximity showed a very strong correlation. The correlation analysis demonstrated that a moderate to a strong statistical dependence between social and mobile proximities exists.

7.1.5 Social Influence

Although the correlation analysis showed a moderate to strong correlation between social and mobile proximities, correlation does not necessarily mean causation. The actual causation effect can be shown by building an influence model for integrating the location histories of friends into the mobility model of an individual. We extended the ST PPM VOMM mobility model for modeling social influences. We refer to the extended model as SOST PPM VOMM. We distinguish between different types of social influences, namely synchronous specific and general (trend) social influences. Synchronous specific social influences have two preconditions, namely the user must be currently involved in a social situation and the location histories of the friends in the current social situation must be available. We categorize the influences arising from the location histories of friends into three classes of synchronous specific social influence factors according to the set of users who build a social influence factor together. In contrast, general social (trends) influences have only one precondition, namely the existence of location histories of friends. The influencers in synchronous specific social influences are the users present in the current social situation of the user, whereas all the friends of the user who have a location history build the influencers in general social (trend) influences. General social trends/influences represent generalities in the mobility of the community of a user, or represent friend recommendations transmitted via unobservable media such as phone, SMS, LBSN platforms (e.g. Foursquare) or mail.

Considering both types of social influence factor, SOST PPM VOMM outperforms ST PPM VOMM by almost 5.22%. The improvement in accuracy occurs mainly during the evening hours of work days and the hours of weekend days starting from 11 a.m. These hours are the typical times when human beings are involved in social activities, thus their mobility exhibits the highest entropy, which means the uncertainty of predicting their next location is at its highest. The incorporation of social networks improves the accuracy of next location prediction during times with high mobility uncertainty, which emphasizes the importance of social influences on the individual mobility of a user. Further, the inclusion of social networks leads to locations which have been previously visited by their friends, but where the user has possibly never been before can be predicted. The improvement in accuracy due to the prediction of locations where the user has never been is 3.19%, which again impressively underlines the importance of social networks.

Similarly to individual location histories, social influences are also subject to decay. We applied two different drift functions, both drift functions led to a significant increase in the accuracy of the model. The degree of drift has shown that social

influences decay within three to six weeks, which is significantly faster than the drift of individual location histories.

Most of the social situations that occurred were between members of maximal 2-plexes. The size of the social situations follows a power law, the majority of the social situations occur between two and five friends and only a few social situations occur between more friends. We calculated a measure of cohesion for each social situation based on the maximal 2-plexes for which the users involved in the social situation are members. The measure of cohesion shows a strong negative correlation with the size of social situations indicating that most cohesive social situations are of a smaller size. The most improvement in accuracy is achieved through social situations among members of maximal 2-plexes. The relationship between the improvement in accuracy and both size and measure of cohesion of the social situation follow power laws. The smaller the size of the social situation, the higher the improvement in accuracy. The higher the measure of cohesion of a social situation, the higher the improvement in accuracy.

Human beings have cognitive, emotional, spatial and temporal limitations, that prevents them from maintaining all their social relationships with the same intensity [Granovetter, 2005]. A user interacts mainly and spends most of their time with their strong ties, who form together the circle of close friends, to which we refered with cohesive subgroups. The emotional needs, beliefs, thoughts, information, locations, times, norms, goals, etc. of the users in the same cohesive subgroup overlap to a high extent. Beside the strong ties, a user has a certain number of social relationships, with whom they interact less, thus, these social relationships represent the weak ties of the user, to which we refered with acquaintances. An interaction between two users connected with a weak tie is most interesting, because the users have their own circle of close friends with own beliefs, thoughts, goals, norms, information, emotional needs, movements, etc. The overlap between their information of both users is considerably less, compared to the members of their own circle of friends. Thus, an interaction between two users connected with a weak tie leads to exchanging more novel information. Centrality is the extent, in which a user is socially connected. A central user has a lot of social ties, most of which are weak ties. Central users have the intrinsic ability of bridging the gap between various social communites, thus they are important for transmitting novel information, trends, influences between different communites. The foursquare dataset contains 22 users with more than 1000 social ties, they represent central users in the social network.

In order to emphasize the importance of weak ties and central users, we set improvement in accuracy and degree in relation. The improvement in accuracy of the 22 central users is lower compared to other users. Generally, the relationship netween both quantities shows a negative tendency. Central users come a lot around and are more explorative in nature, because they have many social relationships and bridge different worlds and communities. Empirical results has shown that the central users visit a lot of unknown locations, and have a larger location history. These results emphasize the importance of central users and weak ties for flow of information (trends) between different communities. Although the mobility of central users can only be improved slightly using their social network, they represent trend setters and transmitters between different social communities and bring about an improvement

in the prediction accuracy for their followers (friends).

We conducted an in-depth correlation analysis of mobility measurement and improvement in accuracy to highlight the importance of social networks in next location prediction. The improvement in accuracy using SOST PPM VOMM compared to ST PPM VOMM shows a strong positive correlation with both entropy and number of locations, whereas it correlates strongly negatively to the average frequency of visit per location. The improvement in accuracy shows only a very slight correlation to the size of location history, because both number of locations and entropy exhibit a positive correlation with the history size. This means a higher history size implies higher entropy and more visited locations. The correlation analysis emphasized the importance of social networks for alleviating the drawbacks of cold-start and zero-frequency problems.

7.1.6 Additional Context & Discrete Knowledge

Context is not limited to spatial, temporal or social features. Today's mobile devices (smart phones) are equipped with a series of sensors which can deliver valuable information about the current context of the user. GSM, W-LAN, Bluetooth, Accelerometer, system programs, alarm clock, etc. are examples of these sensors. Setting the alarm clock may indicate the user is at home or at a hotel. Setting the system profile of the device could indicate the user is entering a meeting room or a doctor's surgery. GSM cell tower IDs or W-LAN access points can be associated with physical locations, Bluetooth observations with other users in the neighborhood of the user, etc. We extended the ST PPM VOMM so that additional information from sources such as the sensors mentioned above could be integrated. We used the Mobile Data Challenge (MDC) [Laurila et al., 2012] dataset to evaluate the extended model. The empirical results show that incorporation of GSM improves the accuracy of ST PMM VOMM from 58.11% to 60.63%. The other sensors improve the prediction accuracy for some users, but on average the improvements are less significant, because many of these sensors depend on the existence of further information. For example, the social network must be known in order to achieve improvements in accuracy while using Bluetooth observations, because observation of other devices in the neighborhood of the user is only relevant when the observed devices belong to friends. Therefore we conclude that these sensors can indeed improve prediction accuracy if they can be combined with other information sources. Unfortunately the MDC dataset does not contain the social network of the user, which would enable us to test the dependency of Bluetooth on the social network of the users.

Human beings might interrupt their daily routine for a visit to the doctor, theatre, cinema, restaurant, cosmetic saloon, music concert, sport stadium, etc. The interruption is temporary and the user soon returns to their normal routine. We refer to knowledge about these visits as discrete knowledge, because there is no pattern to it and it is valid only for a certain period of time. Although these visits occur rarely and irregularly, evidence for many of these visits can be uncovered in advance. For example, users note their appointments in their personal calendar or they dial phone numbers or visit websites corresponding to these locations. Thus they can be detected using information sources like the personal calendar or the web

browser. Although adherence to schedule varies according to the type of appointment and character of the user, users do not generally take the trouble to make an appointment if they do not intend keeping it. A survey has shown that adherence to schedule is very high ($> 90\%$) on average, additionally, people usually cancel when they cannot keep an appointment. We extended ST PPM VOMM to incorporate discrete knowledge by taking user's adherence to schedule into account. Due to the lack of a dataset containing scheduled visits, we enriched the Reality Mining dataset with schedules for around 2% of the visits. The user specific labels for the cell tower IDs allow detection of many visits to doctors, cinemas, theaters, etc. We adjusted the user's adherence to schedule depending on the user's compliance to scheduled visits. The empirical results showed that inclusion of discrete knowledge improves the prediction accuracy by 1.81%.

7.2 Critique

Despite the many positive and innovative aspects our work is not free of critical points. We mention some of these points below, which in our opinion should have received more attention.

The empirical results rely on datasets collected during experiments conducted over a certain time span for a certain number of subjects. The subjects are usually students or university employees. Although the Foursquare dataset contains real-life data, the dataset is very sparse and mainly contains incomplete location histories for the users. Unfortunately, we cannot state whether the same level of accuracy can be achieved for all users, regardless of their occupation or social status. Further, it is possible that the improvements in accuracy achieved by incorporating social networks are in fact due to this artifact of the Foursquare dataset. Figure (4.1) from chapter 4 shows that most check-ins are at publicly accessible locations and only very few check-ins are at significant and frequently visited locations such as home and work locations. The behavior of the model under the availability of complete location histories of the users remains uninvestigated. Hence, we cannot make a statement as to whether the incorporation of social networks would improve the accuracy of the model to the same extent if complete location histories of the users were available. It is possible that the high improvement in accuracy is due to the sparsity of the dataset and the fact it predominantly contains locations where social situations typically evolve. Unfortunately we did not have access to a dataset with complete location histories for the users, but it would be quite feasible to collect such a dataset using mobile devices equipped with GPS, Bluetooth, W-LAN, various Apps and with access to the Internet (like smart phones).

Privacy issues remain a serious issue when predicting the next location of a user. Although next location prediction has many useful applications, it is still a double edged sword and poses risks of privacy violation. People have a basic trust, which as long as they have not yet had a negative experience remains intact. The recent NSA spying scandal [nsa, 2013] has shown the danger of sensitive data getting into the wrong hands, for example, as a vehicle to draw illegal benefits, or in the simplest case for illegal monitoring and surveillance of people. Next location prediction, unfortunately, can also be misused to harm people. It is clear that peoples' trust

in a location-based service can be increased if they have the option to opt-out of the service and delete their data whenever and wherever they want, however an affair like the NSA spying scandal destroys trust in location-based services. Unfortunately we have not paid attention to privacy protection in this work, especially to the possibility of a mobility model algorithm that guarantees an appropriate level of (context-aware) privacy protection.

We believe that the solutions proposed in this work, especially the mobility model, are of low complexity, nevertheless, a further point of criticism is that no attention was paid to the computational complexity (with respect to both time and storage costs) of the proposed solutions. Computational complexity has a direct influence on the architectural design of a mobility model. For example, whether the proposed solutions can run on a mobile device with limited capabilities, or whether client server architecture is needed. The architectural design in turn has influence on the risks of privacy violation and strategies for privacy protection, etc.

The periodic pattern detection method in this work relies on a heuristic a priori approach. The approach exploits naturally occurring periodicities that are valid in Western societies. The approach subdivides the time into weeks, the weeks into work days (Monday-Friday) and weekend (Saturday and Sunday). We assume the universal validity of these periodicities and believe that all users exhibit different mobility behavior on work days compared to their off days. Further, we implicitly assume that all users have their off days on weekend days. The approach might be valid for Western societies, but not for all countries. For example, almost all Islamic societies have Friday off, other countries could have their own specific non-working days. Even in Western societies, not all users have their off days at the weekend, for example the off days of a user may vary according to their profession, most employees in the gastronomy industry have their off days on days other than weekend days. Finally, we haven't considered any country-specific holidays, on which people exhibit a very different mobility behavior than on normal work days. A user and a country-specific periodicity will most probably increase the performance of the proposed mobility model.

The mobility model proposed in this work is able to assign a probability mass to a context (social-spatial-temporal pattern) that does not appear in the location history of the user. We refer to this ability as the escape mechanism. The escape mechanism of the proposed mobility model uses the number of symbols appearing after a context s in the calculation of any probability masses (according to both equation (2.11) and (2.11)). Using the number of symbols appearing after a context for calculating the probability of any context has turned out to be problematic when including the location histories of friends, because the number of symbols appearing after a context increases very quickly, which influences the whole probability distribution. In fact, for this reason we were forced to use a separate tree (SOST PPM VOMM) for bookkeeping influences from the social network. We haven't paid any attention to calculating the escape probabilities based on an alternative approach.

7.3 Outlook

The following fields of study could significantly contribute to improving the task of predicting the next location of a mobile user.

The PPM VOMM models introduced in this work are able to learn and predict simultaneously as has been seen in the previous chapters. Both learning and prediction are based on discrete locations, which means movement-traces in continuous spaces such as GPS, must be transformed into discrete locations in a preprocessing step. A preprocessing step requires separating both learning and prediction steps. But, the PPM VOMM model can be easily combined with an online clusterer such as 2.4.1.2 in order to simultaneously learn and predict in continuous spaces.

Many location-based services use geographical locations without any contextual information about the locations. Geographical locations do not allow a further comparison between locations except using their physical properties such as distance. A semantic location is a form of location enriched with contextual information such as the type of location (restaurant, cinema, etc.), temporal accessibility (open between 10 a.m. and 8 p.m.), the activities that can be performed at these locations (sport, entertainment, eating, drinking, dancing, etc.), specific properties (inexpensive shop, Chinese restaurant, Latin bar, quality goods, all you can eat, etc.). Semantic locations allow a comparison among them, which can help predicting locations where a user has never been before, even without the consideration of location histories of friends. For example if a user often visits Chinese restaurants, then it may be assumed that the user will probably visit a Chinese restaurant for lunch when visiting a city they have never been to before.

Further, semantic locations can bridge the gap between users and locations. The interests of a user can be learned using the semantic properties of the locations they visit. The interests of the users allow comparison between two users even if no spatial overlap exists between their movements. Semantic locations allow the detection of much wider communities than 2-plexes or a circle of friends. Examples of wider communities are the fans of a sports club, a certain singer, a certain life style, etc. These wider communities allow detection of generalities beyond the border of circles of friends. The prediction accuracy of the mobility of a user can then be improved, even if the social network of the user is not given or the location histories of their friends are unknown.

A social tie characterizes the relationship between two socially connected individuals. The strength of a social tie also depends on subjective criteria like emotional intensity and intimacy [Krackhardt, 1992], nevertheless, the strength of a tie can be measured by observable measurements like the amount of information exchanged between the two individuals, the duration of interaction between them and the affection of the individual for the other, for example when the visit of one of the two individuals causes the other individual to visit the same location. The development of powerful mobile devices equipped with many sensors such as smart phones allows the calculation of tie strength based on message exchange (e-mail, SMS, Bluetooth, Instant Messaging, etc.), duration of interaction (phone call), information sharing (file and image exchanges, co-locations, common activities, etc.). Due to the lack of information, SOST PPM VOMM calculates tie strength based on spatial overlap

between the mobility of two friends. A high spatial overlap does not automatically mean a tight relationship between two users. For example, two working colleagues may have a high spatial overlap because they share a frequently visited location, in this case the work location. The strength of a tie in future can be accurately calculated using the above metrics.

Natural language processing (NLP) [nlp, 2013] is concerned with computer-human interactions and the derivation (for example) of information from written and spoken language input. The following tasks of NLP may be of interest in next location prediction. Named Entity Recognition (NER) determines which items in a sequence of text refer to names such as persons, locations or organization names. Natural Language Understanding, which converts a chunk of text into a more formal representation, so it can be further manipulated by computer programs. Speech Recognition, which can determine the textual representation of an audio stream. Information Extraction for extracting semantic information in a text sequence. Rapid developments in NLP in recent years have facilitated the incorporation of more sources of information for retrieving valuable evidence about the mobility of an individual. Examples of new sources of information are, phone calls, SMS, e-mail, messengers (such as Skype), websites, news feeds from local authorities (police), information from local public transport services (public transport timetables), etc. The above sources of information contain a vast amount of valuable information about locations, times, persons, etc. which without doubt can help increase the prediction accuracy of a mobility model.

Bibliography

[Fou, 2012a] (2012a). Foursquare api. https://developer.foursquare.com/, (checked November 2012).

[Fou, 2012b] (2012b). Foursquare categories. http://aboutfoursquare.com/foursquare-categories/, (checked November 2012).

[Eff, 2013] (2013). Activehybrid: Joy knows the way. intelligent energy management. http://www.bmw.com/com/en/insights/technology/efficientdynamics/phase_2/index.html, (checked July 2013).

[Act, 2013] (2013). Activehybrid: Joy knows the way. intelligent energy management. http://www.bmw.com/com/en/insights/technology/efficientdynamics/phase_1/5series_activehybrid_energy_management.html, (checked July 2013).

[Ana, 2013] (2013). Anatomy of facebook. https://www.facebook.com/notes/facebook-data-team/anatomy-of-facebook/10150388519243859, (checked November 2013).

[Nav, 2013] (2013). Autobauer forscht an lernenden navigationssystemen. http://www.elektronikpraxis.vogel.de/hardwareentwicklung/articles/173026/, (checked July 2013).

[wik, 2013a] (2013a). Context awareness. http://en.wikipedia.org/wiki/Context_awareness#cite_note-rosemann2006-1, (checked July 2013).

[cro, 2013] (2013). Crowdsourcing. http://www.merriam-webster.com/dictionary/crowdsourcing, (checked August 2013).

[Dun, 2013] (2013). Dunbar's number). http://en.wikipedia.org/wiki/Dunbar%27s_number, (checked October 2013).

[fac, 2013] (2013). Facebook: Connect with friends and the world around you on facebook. https://www.facebook.com/, (checked April 2013).

[fea, 2013] (2013). Features (pattern recognition). http://en.wikipedia.org/wiki/Features_(pattern_recognition), (checked November 2013).

[fou, 2013] (2013). Foursquare: Find great places near you.. https://foursquare.com/, (checked April 2013).

[Geo, 2013a] (2013a). Geolife: Building social networks using human location history. http://research.microsoft.com/en-us/projects/geolife/, (checked April 2013).

[Geo, 2013b] (2013b). Geolife: Building social networks using human location history. http://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/, (checked April 2013).

[goo, 2013] (2013). Goolge maps geocoding. https://developers.google.com/maps/documentation/geocoding/?hl=en, (checked July 2013).

[Jac, 2013] (2013). Jaccard similarity coefficient. http://en.wikipedia.org/wiki/Jaccard_index, (checked September 2013).

[wik, 2013b] (2013b). Location based services (lbs). http://en.wikipedia.org/wiki/Location-based_service, (checked July 2013).

[mic, 2013] (2013). Location based social networks (lbsn). http://research.microsoft.com/en-us/projects/lbsn/, (checked July 2013).

[nlp, 2013] (2013). Natural language processing (nlp). http://en.wikipedia.org/wiki/Natural_language_processing, (checked October 2013).

[pla, 2013] (2013). Nokia map (former plazes geo-social sn platform). http://maps.nokia.com, (checked April 2013).

[nsa, 2013] (2013). Nsa spying scandal. http://www.spiegel.de/international/topic/nsa_spying_scandal/, (checked October 2013).

[soc, 2013] (2013). Sociometric badge). http://hd.media.mit.edu/badges/, (checked October 2013).

[tip, 2013] (2013). The tipping point. http://en.wikipedia.org/wiki/The_Tipping_Point, (checked October 2013).

[wik, 2013c] (2013c). Trust (social sciences). http://en.wikipedia.org/wiki/Trust_(social_sciences)#cite_note-4, (checked July 2013).

[twi, 2013] (2013). Twitter: Find out what's happening, right now, with the people and organizations you care about. https://twitter.com/, (checked April 2013).

[Abe and Warmuth, 1990] Abe, N. and Warmuth, M. K. (1990). On the computational complexity of approximating distributions by probabilistic automata. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, COLT '90, pages 52–66, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[ACLU, 2010] ACLU (2010). *Location-based Services: Time for a Privacy Check-in.* American Civil Liberties Union (ACLU) of Northern California.

[Adamic and Adar, 2003] Adamic, L. and Adar, E. (2003). Friends and neighbors on the Web. *Social Networks*, 25(3):211–230.

[Agrawal and Srikant, 1995] Agrawal, R. and Srikant, R. (1995). Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, ICDE '95, pages 3–14, Washington, DC, USA. IEEE Computer Society.

[Alba, 1973] Alba, R. (1973). A graph-theoretic definition of a sociometric clique. *Journal of Mathematical Sociology*, 3:113–126.

[Asahara et al., 2011] Asahara, A., Maruyama, K., Sato, A., and Seto, K. (2011). Pedestrian-movement prediction based on mixed markov-chain model. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, pages 25–33, New York, NY, USA. ACM.

[Ashbrook and Starner, 2003] Ashbrook, D. and Starner, T. (2003). Using GPS to learn significant locations and predict movement across multiple users. *Personal Ubiquitous Comput.*, 7(5):275–286.

[Backstrom et al., 2010] Backstrom, L., Sun, E., and Marlow, C. (2010). Find me if you can: Improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 61–70, New York, NY, USA. ACM.

[Bahl and Padmanabhan., 2000] Bahl, P. and Padmanabhan., V. N. (2000). RADAR: an in-building RF-based user location and tracking system. In *Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (Cat. No.00CH37064)*, volume 2, pages 775–784. IEEE.

[Bai et al., 2003] Bai, F., Sadagopan, N., and Helmy, A. (2003). Important: a framework to systematically analyze the impact of mobility on performance of routing protocols for adhoc networks. In *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*, volume 2, pages 825–835.

[Balasundaram. et al., 2009] Balasundaram., B., Butenko, S., Hicks, I. V., and Sachdeva, S. (2009). Clique relaxations in social network analysis:the maximum k-plex problem. In *Operations Research January/February 2011*, volume 59, pages 133–142.

[Bapierre et al., 2011] Bapierre, H., Groh, G., and Theiner, S. (2011). A variable order markov model approach for mobility prediction. *STAMI2011@IJCAI2011, Barcelona, Spain.*, pages 661–703.

[Barnes and Scornavacca, 2004] Barnes, S. J. and Scornavacca, E. (2004). Mobile marketing: the role of permission and acceptance. *Int. J. Mob. Commun. Inderscience Publishers.*, 2(2):128–139.

[Barwise and Strong, 2002] Barwise, P. and Strong, C. (2002). Permission-based mobile advertising. *Journal of Interactive Marketing*, 16(1"):14 – 24.

[Baum et al., 1970] Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics. Institute of Mathematical Statistics.*, 41(1):164–171.

[Begleiter et al., 2004] Begleiter, R., El-yaniv, R., and Yona, G. (2004). On prediction using variable order markov models. *Journal of Artificial Intelligence Research*, 22:385–421.

[Bettstetter, 2001a] Bettstetter, C. (2001a). Mobility modeling in wireless networks: Categorization, smooth movement, and border effects. In *in ACM Mobile Computing and Communications Review*, volume 5, pages 55–67.

[Bettstetter, 2001b] Bettstetter, C. (2001b). Smooth is better than sharp: A random mobility model for simulation of wireless networks. In *in Proc. ACM Intern. Workshop on Modeling, Analysis, and Simulation of Wireless and Mobile Systems (MSWiM), Rome, Italy, July 2001.*

[Bishop, 2007] Bishop, C. M. (2007). *Pattern Recongnition and Machine Learning.* Springer Science + Business Media LLC., second edition.

[Brown et al., 2011] Brown, D., Farrier, D., Egger, S., McNamara, L., Grewcock, M., and Spears, D. (2011). *Criminal Laws: Materials and Commentary on Criminal Law and Process in New South Wales.* Federation Press.

[Camp et al., 2002] Camp, T., Boleng, J., and Davies, V. (2002). A Survey of Mobility Models for Ad Hoc Network Research. *Wireless Communications & Mobile Computing (WCMC): Special issue on Mobile Ad Hoc Networking: Research, Trends and Applications*, 2(5):483–502.

[Cho et al., 2011] Cho, E., Myers, S. A., and Leskovec, J. (2011). Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 1082–1090, New York, NY, USA. ACM.

[Chon et al., 2012] Chon, Y., Shin, H., Talipov, E., and Cha, H. (2012). Evaluating mobility models for temporal prediction with high-granularity mobility data. IEEE International Conference on Pervasive Computing and Communications (PerCom 2012) Lugano, Switzerland.

[Cormen et al., 2001] Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2001). *Introduction to Algorithms, Second Edition.* The MIT Press, 2 edition.

[Cormen et al., 2009] Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009). *Introduction to Algorithms.* MIT Press, 3rd edition.

[Crandall et al., 2010] Crandall, D. J., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D., and Kleinberg, J. (2010). Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences.*

[Cranshaw et al., 2010] Cranshaw, J., Toch, E., Hong, J., Kittur, A., and Sadeh, N. (2010). Bridging the gap between physical location and online social networks. In the Proceedings of the Twelfth International Conference on Ubiquitous Computing.

[Dey, 2001] Dey, A. K. (2001). Understanding and using context. *Personal Ubiquitous Comput. Springer-Verlag.*, 5(1):4–7.

[Dietterich, 2002] Dietterich, T. G. (2002). Machine learning for sequential data: A review. In *Structural, Syntactic, and Statistical Pattern Recognition*, pages 15–30. Springer-Verlag.

[Dong et al., 2011] Dong, W., Lepri, B., and Pentland, A. S. (2011). Modeling the co-evolution of behaviors and social relationships using mobile phone data. In *Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia*, MUM '11, pages 134–143, New York, NY, USA. ACM.

[Duda et al., 2001] Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. John Wiley and Sons, Inc., second edition.

[Dunbar, 1992] Dunbar, R. I. M. (1992). Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22(6):469–493.

[Eagle et al., 2007] Eagle, N., Pentland, A., and Lazer, D. (2007). Inferring social network structure using mobile phone data. *PNAS*.

[Eagle and Pentland, 2006] Eagle, N. and Pentland, A. S. (2006). Reality mining: sensing complex social systems. *Journal Personal and Ubiquitous Computing Volume 10 Issue 4, March 2006 Pages 255 - 268*, 10(1):255–268.

[Eagle and Pentland, 2009] Eagle, N. and Pentland, A. S. (2009). Eigenbehaviors: identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066.

[Elliott and Stettler, 2007] Elliott, M. R. and Stettler, N. (2007). Using a mixture model for multiple imputation in the presence of outliers: the healthy for life project. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 56(1):63–78.

[Ericsson et al., 2006] Ericsson, E., Larsson, H., and Brundell-Freij, K. (2006). Optimizing route choice for lowest fuel consumption - potential effects of a new driver support tool. *Transportation Research Part C: Emerging Technologies*, 14(6):369 – 383.

[Ester et al., 1996] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise*. In Evangelos Simoudis, Jiawei Han, Usama M. Fayyad. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press.

[Etter et al., 2012] Etter, V., Kafsi, M., and Kazemi, E. (2012). Been there, done that: What your mobility traces reveal about your behavior. *the Procedings of Mobile Data Challenge by Nokia Workshop at the Tenth International Conference on Pervasive Computing.*

[Everett, 1982] Everett, M. G. (1982). Graph theoretic blockings k-plexes and k-cutpoints. *Journal of Mathematical Sociology*, 9:75–84.

[Fano, 1961] Fano, R. (1961). *Transmission of Information: A Statistical Theory of Communications.* MIT Press Classics. M.I.T. Press.

[Festinger et al., 1950] Festinger, L., Schachter, S., and Back, K. (1950). *The Spatial Ecology of Group Formation.* in L. Festinger, S. Schachter, & K. Back (eds.), Social Pressure in Informal Groups, 1950. Chapter 4.

[Fischer and Wiswede, 1997] Fischer, L. and Wiswede, G. (1997). *Grundlagen der Sozialpsychologie.* Oldenbourg Verlag.

[Flynn et al., 2010] Flynn, F., Reagans, R., and Guillory, L. (2010). Do you two know each other? transitivity, homophily, and the need for (network) closure. *Journal of Personality and Social Psychology.*, 99(5):855–869.

[Ganti et al., 2010] Ganti, R. K., Pham, N., Ahmadi, H., Nangia, S., and Abdelzaher, T. F. (2010). Greengps: a participatory sensing fuel-efficient maps application. In *Proceedings of the 8th international conference on Mobile systems, applications, and services*, MobiSys '10, pages 151–164, New York, NY, USA. ACM.

[Gao et al., 2011a] Gao, H., Barbier, G., and Goolsby, R. (2011a). Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems. IEEE Educational Activities Department.*, 26(3):10–14.

[Gao et al., 2012] Gao, H., Tang, J., and Liu, H. (2012). Mobile location prediction in spatio-temporal context. *the Procedings of Mobile Data Challenge by Nokia Workshop at the Tenth International Conference on Pervasive Computing. Arizona State University.*

[Gao et al., 2011b] Gao, H., Wang, X., Barbier, G., and Liu, H. (2011b). Promoting coordination for disaster relief: from crowdsourcing to coordination. In *Proceedings of the 4th international conference on Social computing, behavioral-cultural modeling and prediction*, SBP'11, pages 197–204, Berlin, Heidelberg. Springer-Verlag.

[Gedik and Liu, 2005] Gedik, B. and Liu, L. (2005). Location privacy in mobile systems: A personalized anonymization model. In *Distributed Computing Systems, 2005. ICDCS 2005. Proceedings. 25th IEEE International Conference on*, pages 620–629.

[Giannotti et al., 2006] Giannotti, F., Nanni, M., and Pedreschi, D. (2006). Efficient mining of temporally annotated sequences. In Ghosh, J., Lambert, D., Skillicorn, D. B., and Srivastava, J., editors, *SDM*. SIAM.

[Gladwell, 2002] Gladwell, M. (2002). *The Tipping Point: How Little Things Can Make a Big Difference*. Back Bay Books.

[Goldenberg and Levy, 2009] Goldenberg, J. and Levy, M. (2009). Distance is not dead: Social interaction and geographical distance in the internet era. *arXiv preprint arXiv:0906.3202*.

[Gong et al., 2010] Gong, Z., Sun, G.-Z., and Xie, X. (2010). Protecting privacy in location-based services using k-anonymity without cloaked region. *Mobile Data Management, IEEE International Conference on. IEEE Computer Society.*, 0:366–371.

[Gonzalez et al., 2008] Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature. Nature Publishing Group*, 453(7196):779–782.

[Gonzalez et al., 2006] Gonzalez, M. C., Lind, P. G., and Herrmann, H. J. (2006). A system of mobile agents to model social networks. *Phys. Rev. Lett. American Physical Society.*

[Goodman, 1999] Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The p value fallacy. *Annals of internal medicine. American College of Physicians.*, 130(12):995–1004.

[Granovetter, 1982] Granovetter, M. (1982). The strength of weak ties: A network theory revisited. *Sociological Theory*, pages 105–130.

[Granovetter, 2005] Granovetter, M. (2005). The impact of social structure on economic outcomes. *The Journal of Economic Perspectives*, 19(1):33–50.

[Groh, 2005] Groh, G. (2005). *Ad-Hoc-Groups in Mobile Communities: Detection, Modeling and Applications.* PhD thesis.

[Groh et al., 2010] Groh, G., Lehmann, A., Reimers, J., Friess, M. R., and Schwarz, L. (2010). Detecting social situations from interaction geometry. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing*, SOCIALCOM '10, pages 1–8, Washington, DC, USA. IEEE Computer Society.

[Gruteser and Grunwald, 2003] Gruteser, M. and Grunwald, D. (2003). Anonymous usage of location-based services through spatial and temporal cloaking. In *Proceedings of the 1st international conference on Mobile systems, applications and services*, MobiSys '03, pages 31–42, New York, NY, USA. ACM.

[Gupta et al., 2009] Gupta, M., Intille, S. S., and Larson, K. (2009). Adding gps-control to traditional thermostats: An exploration of potential energy savings and design challenges. In *Proceedings of the 7th International Conference on Pervasive Computing*, Pervasive '09, pages 95–114, Berlin, Heidelberg. Springer-Verlag.

[Gustafsson et al., 2002] Gustafsson, F., Gunnarsson, F., Bergman, N., Forssell, U., Jansson, J., Karlsson, R., and Nordlund, P. J. (2002). Particle filters for positioning, navigation, and tracking. *IEEE Transactions on Signal Processing. IEEE.*, 50(2):425–437.

[Haas, 1997] Haas, Z. (1997). A new routing protocol for reconfigurable wireless networks. pages 562–565. In Proceedings of the IEEE International Conference on Universal Personal Communications (ICUPC).

[Hackney and Axhausen, 2006] Hackney, J. and Axhausen, K. W. (2006). Agent model of social network and travel behavior interdependence. *paper presented at the 11th International Conference on Travel Behaviour Research Kyoto*, pages 16–20.

[Han et al., 2007] Han, J., Cheng, H., Xin, D., and Yan, X. (2007). Frequent pattern mining: current status and future directions. *Data Min. Knowl. Discov.*, 15(1):55–86.

[Han et al., 1999] Han, J., Dong, G., and Yin, Y. (1999). Efficient mining of partial periodic patterns in time series database. In *Data Engineering, 1999. Proceedings., 15th International Conference on*, pages 106–115.

[Han et al., 1998] Han, J., Gong, W., and Yin, Y. (1998). Mining segment-wise periodic patterns in time-related databases. In *Proc. Int. Conf. on Knowledge Discovery and Data Mining*, pages 214–218.

[Hanneman and Riddle, 2005] Hanneman, R. A. and Riddle, M. (2005). Introduction to social network methods. http://faculty.ucr.edu/~hanneman/nettext/C11_Cliques.html, (checked March 2013).

[Hanneman and Riddle, 2011] Hanneman, R. A. and Riddle, M. (2011). *Concepts and Measures for Basic Network Analysis.* The Sage Handbook of Social Network Analysis.

[Homans, 2001] Homans, G. (2001). *The Human Group.* Reprint Transaction Pub., New York, 2001 (org. 1950).

[Hong and Rappaport Stephen, 1986] Hong, D. and Rappaport Stephen, S. (1986). Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures. *Vehicular Technology, IEEE Transactions on*, 35(3):77–92.

[Hong et al., 1999] Hong, X., Gerla, M., Pei, G., and Chiang, C. (1999). A group mobility model for ad hoc wireless networks. In Proceedings of the ACM International Workshop on Modeling and Simulation of Wireless and Mobile Systems (MSWiM).

[Horton and Lipsitz, 2001] Horton, N. J. and Lipsitz, S. R. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *The American Statistician*, 55(3):244–254.

[Howe, 2006] Howe, J. (2006). The rise of crowdsourcing. *Wired magazine*, 14(6):1–4.

[Jardosh et al., 2003] Jardosh, A., BeldingRoyer, E. M., Almeroth, K. C., and Suri, S. (2003). Towards realistic mobility models for mobile ad hoc networks. pages

217–229. In Proceeding MobiCom '03 Proceedings of the 9th annual international conference on Mobile computing and networking.

[Jesdabodi, 2012] Jesdabodi, C. (2012). Modeling the influence of the social network on the location behavior of users. Master's thesis, Fakultät für Informatik der Technischen Universität München. Supervised by Georg Groh and Halgurt Bapierre.

[Jiang et al., 2010] Jiang, J., Wilson, C., Wang, X., Huang, P., Sha, W., Dai, Y., and Zhao, B. Y. (2010). Understanding latent interactions in online social networks. In Proceedings of IMC 10. ACM, 2010.

[Johnson and Maltz, 1996] Johnson, D. B. and Maltz, D. A. (1996). Dynamic source routing in ad hoc wireless networks. In *Mobile Computing*, pages 153–181. Kluwer Academic Publishers.

[Kadushin, 2012] Kadushin, C. (2012). Understanding social networks: Theories, concepts, and findings. *Oxford University Press.*

[Kaemarungsi, 2005] Kaemarungsi, K. (2005). Efficient design of indoor positioning systems based on location fingerprinting. In *Wireless Networks, Communications and Mobile Computing, 2005 International Conference on*, volume 1, pages 181–186 vol.1.

[Kaemarungsi and Krishnamurthy, 2004] Kaemarungsi, K. and Krishnamurthy, P. (2004). Modeling of indoor positioning systems based on location fingerprinting. In *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies*, volume 2, pages 1012–1022.

[Kalnis et al., 2007] Kalnis, P., Ghinita, G., Mouratidis, K., and Papadias, D. (2007). Preventing location-based identity inference in anonymous spatial queries. *Knowledge and Data Engineering, IEEE Transactions on*, 19(12):1719–1733.

[Kaltenbrunner et al., 2012] Kaltenbrunner, A., Scellato, S., Volkovich, Y., Laniado, D., Currie, D., Jutemar, E. J., and Mascolo, C. (2012). Far from the eyes, close on the web: impact of geographic distance on online social interactions. In the Proceeding of ACM SIGCOMM Workshop on Online Social Networks.

[Kang et al., 2004] Kang, J. H., Welbourne, W., Stewart, B., and Borriello, G. (2004). Extracting places from traces of locations. In *Proceedings of the 2nd ACM international workshop on Wireless mobile applications and services on WLAN hotspots*, WMASH '04, pages 110–118, New York, NY, USA. ACM.

[Kang et al., 2009] Kang, S.-Y., Song, J.-W., Lee, K.-J., Lee, J.-H., Kim, J.-H., and Yang, S.-B. (2009). Improved location acquisition algorithms for the location-based alert service. In *Proceedings of the 3rd International Conference and Workshops on Advances in Information Security and Assurance*, ISA '09, pages 461–470, Berlin, Heidelberg. Springer-Verlag.

[Katz, 1953] Katz, L. (1953). *A new status index derived from sociometric analysis.*, volume 18(1). Psychometrika.

[Kim and Kotz, 2011] Kim, M. and Kotz, D. (2011). Identifying unusual days. *Journal of Computing Science and Engineering (JCSE)*.

[Kosub, 2004] Kosub, S. (2004). Local density. network analysis; springer lncs 3418. Master's thesis.

[Krackhardt, 1992] Krackhardt, D. (1992). The strength of strong ties: The importance of philos in organizations. *Harvard Business School Press*, pages 216–239.

[Krumm and Brush, 2011] Krumm, J. and Brush, A. J. B. (2011). Learning time-based presence probabilities. In *Proceedings of the 9th international conference on Pervasive computing*, Pervasive'11, pages 79–96, Berlin, Heidelberg. Springer-Verlag.

[Krumm and Horvitz, 2004] Krumm, J. and Horvitz, E. (2004). Locadio: Inferring motion and location from wi-fi signal strengths. In *in First Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services (Mobiquitous*.

[Krumm and Horvitz, 2006] Krumm, J. and Horvitz, E. (2006). Predestination: Inferring destinations from partial trajectories. In *Proceedings of the 8th International Conference on Ubiquitous Computing*, UbiComp'06, pages 243–260, Berlin, Heidelberg. Springer-Verlag.

[Krumm et al., 2003] Krumm, J., Scellato, S., Laniado, D., Mascolo, C., and Kaltenbrunner, A. (2003). Probabilistic inferencing for location. In *Microsoft Research, Microsoft Corporation, One Microsoft Way, Redmond, WA USA, jckrumm@microsoft.com*. 2003 Workshop on Location-Aware Computing (Part of UbiComp 2003), Seattle, WA, USA.

[Kschischang et al., 2001] Kschischang, F., Frey, B., and Loeliger, H.-A. (2001). Factor graphs and the sum-product algorithm. *Information Theory, IEEE Transactions on*, 47(2):498–519.

[Küpper, 2005] Küpper, A. (2005). *Location-Based Services: Fundamentals and Operation*. John Wiley & Sons.

[Lambiotte et al., 2008] Lambiotte, R., Blondel, V. D., de Kerchove, C., Huens, E., Prieur, C., Smoreda, Z., and Dooren, P. V. (2008). Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325.

[Laurila et al., 2012] Laurila, J. K., Gatica-Perez, D., Aad, I., J., B., Bornet, O., Do, T.-M.-T., Dousse, O., Eberle, J., and Miettinen, M. (2012). The mobile data challenge: Big data for mobile computing research. In *Pervasive Computing*.

[Lee et al., 2011] Lee, K. H., Lippman, A., Pentland, A., and Dugundji, E. (2011). The impacts of just-in-time social networks on people's choices in the real world. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 9–18.

[Lee et al., 2008] Lee, S. H., Walters, S. D., and Howlett, R. J. (2008). Intelligent gps-based vehicle control for improved fuel consumption and reduced emissions. In *Proceedings of the 12th international conference on Knowledge-Based Intelligent Information and Engineering Systems, Part III*, KES '08, pages 701–708, Berlin, Heidelberg. Springer-Verlag.

[Lehmann, 2010] Lehmann, A. (2010). Towards mobile location- and orientation-based detection of social situations. diploma thesis, tu-münchen, ws 2009 / 2010; supervisor: Georg groh. Master's thesis.

[Li et al., 2008] Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., and Ma, W.-Y. (2008). Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, GIS '08, pages 34:1–34:10, New York, NY, USA. ACM.

[Liang and Haas, 2003] Liang, B. and Haas, Z. J. (2003). Predictive distance-based mobility management for multidimensional PCS networks. *IEEE/ACM Trans. Netw. IEEE Press.*, 11(5):718–732.

[Liao et al., 2003] Liao, L., Fox, D., Hightower, J., Kautz, H., and Schulz, D. (2003). Voronoi tracking: location estimation using sparse and noisy sensor data. In *Intelligent Robots and Systems, 2003. (IROS 2003). Proceedings. 2003 IEEE/RSJ International Conference on*, volume 1, pages 723–728 vol.1.

[Liao et al., 2007a] Liao, L., Fox, D., and Kautz, H. (2007a). Extracting places and activities from gps traces using hierarchical conditional random fields. *Int. J. Rob. Res. Sage Publications, Inc.*, 26(1):119–134.

[LIAO et al., 2006] LIAO, L., PATTERSON, D. J., FOX, D., and KAUTZ, H. (2006). Building personal maps from gps data. *Annals of the New York Academy of Sciences. Blackwell Publishing Inc.*, 1093(1):249–265.

[Liao et al., 2007b] Liao, L., Patterson, D. J., Fox, D., and Kautz, H. (2007b). Learning and inferring transportation routines. *Artif. Intell. Elsevier Science Publishers Ltd.*, 171(5-6):311–331.

[Liben-Nowell and Kleinberg, 2003] Liben-Nowell, D. and Kleinberg, J. (2003). The link prediction problem for social networks. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, pages 556–559, New York, NY, USA. ACM.

[Little and Rubin, 2002] Little, R. and Rubin, D. (2002). *Statistical analysis with missing data*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley, Hoboken, New Jersey.

[Little, 1988] Little, R. J. A. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6(3):287–96.

[Lu and Liu, 2012] Lu, Y. and Liu, Y. (2012). Pervasive location acquisition technologies: Opportunities and challenges for geospatial studies. *Computers, Environment and Urban Systems*, pages 105–108.

[Luce, 1950] Luce, R. (1950). Connectivity and generalized cliques in sociometric group structure. *Psychometrika*, 15:169–190.

[Luce and Perry, 1949] Luce, R. and Perry, A. (1949). A method of matrix analysis of group structure. *Psychometrika*, 14(2):95–116.

[Lucille and Powell, 1975] Lucille, N. and Powell, L. M. (1975). Similarity and propinquity in friendship formation. *Journal of Personality and Social Psychology.*, 32(2):205–213.

[Mackie, 1974] Mackie, J. L. (1974). *The Cement of the Universe: A Study of Causation.* Clarendon Press, Oxford.

[Makino and Uno, 2004] Makino, K. and Uno, T. (2004). New algorithms for enumerating all maximal cliques. pages 260–272. Springer-Verlag.

[Mathew et al., 2012] Mathew, W., Raposo, R., and Martins, B. (2012). Predicting future locations with hidden markov models. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 911–918, New York, NY, USA. ACM.

[Mcfadden, 2001] Mcfadden, D. (2001). Economic choices. *American Economic Review*, 91:351–378.

[McPherson et al., 2001] McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). *Birds of a Feather: Homophily in Social Networks.* Annual Reviews.

[Milgram, 1967] Milgram, S. (1967). The Small World Problem. *Psychology Today*, 2:60–67.

[Mokken, 1979] Mokken, R. J. (1979). Cliques, clubs and clans. *Quality & Quantity. Springer Netherlands*, 13(2):161–173.

[Mooney and Roddick, 2013] Mooney, C. H. and Roddick, J. F. (2013). Sequential pattern mining – approaches and algorithms. *ACM Comput. Surv. ACM.*, 45(2):19:1–19:39.

[Moreno, 1951] Moreno, J. L. (1951). Sociometry, experimental method and the science of society. an approach to a new political orientation. *Beacon House, Beacon, New York.*

[Musolesi et al., 2004] Musolesi, M., Hailes, S., and Mascolo, C. (2004). An ad hoc mobility model founded on social network theory. In *Proceedings of the 7th ACM International Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, MSWiM '04, pages 20–24, New York, NY, USA. ACM.

[Newman, 2001] Newman, M. (2001). Clustering and preferential attachment in growing networks. *Physical Review E. APS.*, 64(2):025102.

[Newman, 2003] Newman, M. E. J. (2003). The Structure and Function of Complex Networks. *SIAM Review*, 45(2):167–256.

[Noulas et al., 2012] Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., and Mascolo, C. (2012). A tale of many cities: Universal patterns in human urban mobility. *PLoS ONE. Public Library of Science.*, 7(5):e37027.

[Noulas et al., 2011a] Noulas, A., Scellato, S., Mascolo, C., and Pontil, M. (2011a). An empirical study of geographic user activity patterns in foursquare. In *Proc. of the 5th Int'l AAAI Conference on Weblogs and Social Media*, pages 570–573.

[Noulas et al., 2011b] Noulas, A., Scellato, S., Mascolo, C., and Pontil, M. (2011b). Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. In *The Social Mobile Web*, volume WS-11-02 of *AAAI Workshops*. AAAI.

[Nowell et al., 2005] Nowell, D. L., Novak, J., Kumar, R., Raghavan, P., Tomkins, A., and Graham, R. L. (2005). Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623–11628.

[Onnela et al., 2007] Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., and Barabási, A.-L. (2007). Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336.

[Pan et al., 2012] Pan, W., Dong, W., Cebrian, M., Kim, T., Fowler, J., and Pentland, A. (2012). Modeling dynamical influence in human interaction: Using data to make better inferences about influence within social systems. *Signal Processing Magazine, IEEE*, 29(2):77–86.

[Patterson et al., 2003a] Patterson, A., Muntz, R. R., and Pancake, C. M. (2003a). Challenges in location-aware computing. *IEEE Pervasive Computing*, 2:80–89.

[Patterson et al., 2002] Patterson, D. J., Etzioni, O., Fox, D., and Kautz, H. (2002). Intelligent ubiquitous computing to support alzheimer's patients: Enabling the cognitively disabled. In *In UbiCog '02: First International Workshop on Ubiquitous Computing for Cognitive Aids*.

[Patterson et al., 2003b] Patterson, D. J., Liao, L., Fox, D., and Kautz, H. (2003b). Inferring high-level behavior from low-level sensors. In *In International Conference on Ubiquitous Computing*, pages 73–89.

[Patterson et al., 2004] Patterson, D. J., Liao, L., Gajos, K., Collier, M., Livic, N., Olson, K., Wang, S., Fox, D., and Kautz, H. (2004). Opportunity knocks: a system to provide cognitive assistance with transportation services. In *In International Conference on Ubiquitous Computing (UbiComp*, pages 433–450. Springer.

[Perusco and Michael, 2007] Perusco, L. and Michael, K. (2007). Control, trust, privacy, and security: evaluating location-based services. *IEEE Technology and Society Magazine*, 26(1):4–16.

[Pesaran et al., 2003] Pesaran, A., Vlahinos, A., and Stuart, T. (2003). Cooling and preheating of batteries in hybrid electric vehicles. In *6th ASME-JSME Thermal Engineering Joint Conference*.

[Petzold et al., 2005] Petzold, J., Pietzowski, A., Bagci, F., Trumler, W., and Ungerer, T. (2005). Prediction of indoor movements using bayesian networks. In *In Proceedings of Location- and Context-Awareness (LoCA 2005.*

[Poolsappasit and Ray, 2009] Poolsappasit, N. and Ray, I. (2009). Towards achieving personalized privacy for location-based services. *Trans. Data Privacy. IIIA-CSIC.*, 2(1):77–99.

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

[Raghunathan et al., 2001] Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology*, 27(1):85–96.

[Rasheed, 2011] Rasheed, F. (2011). *Efficient periodic pattern mining in time series & sequence databases.* PhD thesis, Calgary, Alta., Canada, Canada. AAINR75499.

[Riegelman, 1979] Riegelman, R. (1979). Contributory cause: unnecessary and insufficient. *Postgrad Med*, 66(2):177–9.

[Rodriguez-Carrion et al., 2012] Rodriguez-Carrion, A., Garcia-Rubio, C., Campo, C., Cortis-Martin, A., Garcia-Lozano, E., and Noriega-Vivas, P. (2012). Study of lz-based location prediction and its application to transportation recommender systems. *Sensors*, 12(6):7496–7517.

[Ron and Singer, 1996] Ron, D. and Singer, Y. (1996). The power of amnesia: learning probabilistic automata with variable memory length. In *Machine Learning*, volume 25, pages 117–149.

[Rubin and Schenker, 1986] Rubin, D. B. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81:366–374.

[Russell and Norvig, 2010] Russell, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach.* Pearson Education, Inc., publiching as Prentice Hall, Upper Saddle River, New Jersey 07458., third edition.

[Sadilek et al., 2012] Sadilek, A., Kautz, H., and Bigham, J. P. (2012). Finding your friends and following them to where you are. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 723–732, New York, NY, USA. ACM.

[Sánchez and Manzoni, 2001] Sánchez, M. and Manzoni, P. (2001). Anejos: A java based simulator for ad hoc networks. *Future Gener. Comput. Syst. Elsevier Science Publishers B. V.*, 17(5):573–583.

[Scellato et al., 2011a] Scellato, S., Musolesi, M., Mascolo, C., Latora, V., and Campbell, A. (2011a). NextPlace: A spatio-temporal prediction framework for pervasive systems. In Lyons, K., Hightower, J., and Huang, E., editors, *Pervasive Computing*, volume 6696 of *Lecture Notes in Computer Science*, chapter 10, pages 152–169. Springer Berlin / Heidelberg, Berlin, Heidelberg.

[Scellato et al., 2011b] Scellato, S., Noulas, A., Lambiotte, R., and Mascolo, C. (2011b). Socio-spatial properties of online location-based social networks. In Adamic, L. A., Baeza-Yates, R. A., and Counts, S., editors, *ICWSM*. The AAAI Press.

[Scellato et al., 2011c] Scellato, S., Noulas, A., and Mascolo, C. (2011c). Exploiting place features in link prediction on location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1046–1054, New York, NY, USA. ACM.

[Schmidt, 2002] Schmidt, A. (2002). *Ubiquitous Computing - Computing in Context*. PhD thesis, Lancaster University.

[Schmidt et al., 1999] Schmidt, A., Beigl, M., and Gellersen, H. (1999). There is more to context than location. *Computers & Graphics*, 23(6):893–901.

[Seidman and Foster, 1978] Seidman, S. B. and Foster, B. L. (1978). A graph theoretic generalization of the clique concept. *Journal of Mathematical Sociology*, 6:139–154.

[Sheng et al., 2006] Sheng, C., Hsu, W., and Lee, M. L. (2006). Mining dense periodic patterns in time series data. In *Proceedings of the 22nd International Conference on Data Engineering*, ICDE '06, pages 115–, Washington, DC, USA. IEEE Computer Society.

[Shin et al., 2012] Shin, K. G., Ju, X., Chen, Z., and Hu, X. (2012). Privacy protection for users of location-based services. *IEEE Wireless Commun.*, 19(2):30–39.

[Song et al., 2010a] Song, C., Koren, T., Wang, P., and Barabasi, A.-L. (2010a). Modelling the scaling properties of human mobility. *Nature Physics. Nature Publishing Group.*, 6(10):818–823.

[Song et al., 2010b] Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. (2010b). Limits of predictability in human mobility. *Science. American Association for the Advancement of Science.*, 327(5968):1018–1021.

[Srikant and Agrawal, 1996] Srikant, R. and Agrawal, R. (1996). Mining sequential patterns: Generalizations and performance improvements. In *Proceedings of the 5th International Conference on Extending Database Technology: Advances in Database Technology*, EDBT '96, pages 3–17, London, UK, UK. Springer-Verlag.

[Tang et al., 2009] Tang, J., Sun, J., Wang, C., and Yang, Z. (2009). Social influence analysis in large-scale networks. *KDD 2009 Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 807–816.

[Thomas and of Sociology, 2009] Thomas, R. J. and of Sociology, S. U. D. (2009). *Geographic mobility and homophily*. Stanford University.

[Timm et al., 2003] Timm, H., Döring, C., and Kruse, R. (2003). Using association rules for completing missing data. *International Journal of Approximate Reasoning, 35*.

[Tobler, 1970] Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. volume 46, pages 234–240. Economic Geography.

[Tran et al., 2008] Tran, K., Phung, D., Adams, B., and Venkatesh, S. (2008). Indoor location prediction using multiple wireless received signal strengths. In *Proceedings of the 7th Australasian Data Mining Conference - Volume 87*, AusDM '08, pages 187–192, Darlinghurst, Australia, Australia. Australian Computer Society, Inc.

[Tran et al., 2012] Tran, L. H., Catasta, M., McDowell, L. K., and Aberer, K. (2012). Next Place Prediction using Mobile Data. In *Proceedings of the Mobile Data Challenge Workshop (MDC 2012)*.

[Travers and Milgram, 1969] Travers, J. and Milgram, S. (1969). An Experimental Study of the Small World Problem. *Sociometry. American Sociological Association.*, 32(4):425–443.

[Tsukiyama et al., 1977] Tsukiyama, S., Ide, M., Ariyoshi, H., and Shirakawa, I. (1977). A new algorithm for generating all the maximal independent sets. *SIAM Journal on Computing*, 6(3):505–517.

[U.Brandes and Erlebach, 2004] U.Brandes and Erlebach, T. (2004). *"Fundamentals" in U.Brandes, T. Erlebach (Eds.): Network Analysis;*. Springer LNCS 3418.

[Varshney, 2001] Varshney, U. (2001). Location management support for mobile commerce applications. *International Conference on Mobile Computing and Networking*, pages 1–6.

[Vintan et al., 2004] Vintan, L., Gellert, A., Petzold, J., and Ungerer, T. (2004). Person movement prediction using neural networks. In *In First Workshop on Modeling and Retrieval of Context*.

[Volkovich et al., 2012] Volkovich, Y., Scellato, S., Laniado, D., Mascolo, C., and Kaltenbrunner, A. (2012). The length of bridge ties: structural and geographic properties of online social interactions. In Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media.

[Vrček et al., 2009] Vrček, N., Bubaš, G., and Bosilj, N. (2009). User acceptance of location-based services. *International Journal of Human and Social Sciences*, 4:2.

[Vu et al., 2011] Vu, L., Do, Q., and Nahrstedt, K. (2011). Jyotish: A novel framework for constructing predictive model of people movement from joint wifi/bluetooth trace. In *PerCom*, pages 54–62. IEEE.

[Wakita and Tsurumi, 2007] Wakita, K. and Tsurumi, T. (2007). Finding community structure in megascale social networks. *In Proceedings of the 16th international conference on World Wide Web Pages 1275-1276*.

[Wang et al., 2011] Wang, D., Pedreschi, D., Song, C., Giannotti, F., and Barabasi, A. L. (2011). Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 1100–1108, New York, NY, USA. ACM.

[Wang and Prabhala, 2012] Wang, J. and Prabhala, B. (2012). Periodicity based next place prediction. *the Procedings of Mobile Data Challenge by Nokia Workshop at the Tenth International Conference on Pervasive Computing. University of Illinois at Urbana.*

[Wasserman and Faust, 1994] Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications.* Cambridge: Cambridge University Press.

[Watts, 2004] Watts, D. J. (2004). *Six Degrees: The Science of a Connected Age.* W. W. Norton & Company.

[Watts and Strogatz, 1998] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, Volume 393, Issue 6684)*, 393:440–442.

[Wevers et al., 2010] Wevers, K., Loewenau, J., Durekovic, S., and Lu, M. (2010). interactive-high precision maps for sustainable accident reduction with the enhanced dynamic pass predictor.

[Wilson et al., 2009] Wilson, C., Boe, B., Sala, A., Puttaswamy, K. P. N., and Zhao, B. Y. (2009). User interactions in social networks and their implications. In Proceedings of EuroSys 09. ACM, 2009.

[Wu et al., 2004] Wu, C.-H., Wun, C.-H., and Chou, H.-J. (2004). Using association rules for completing missing data. In *Hybrid Intelligent Systems, 2004. HIS '04. Fourth International Conference on*, pages 236–241.

[Xu et al., 2010] Xu, G., Zhang, Y., and Li, L. (2010). *Web Mining and Social Networking: Techniques and Applications.* Springer-Verlag New York, Inc., New York, NY, USA, 1st edition.

[Ye et al., 2009] Ye, Y., Zheng, Y., Chen, Y., Feng, J., and Xie, X. (2009). Mining individual life pattern based on location history. In *Proceedings of the 2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware*, MDM '09, pages 1–10, Washington, DC, USA. IEEE Computer Society.

[Yiu et al., 2008] Yiu, M. L., Jensen, C. S., Huang, X., and Lu, H. (2008). Spacetwist: Managing the trade-offs among location privacy, query performance, and query accuracy in mobile services. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pages 366–375.

[Yu et al., 2006] Yu, X., Liu, Y., Wei, D., and yin Lei, L. (2006). A hybrid markov model based on em algorithm. In *ICARCV*, pages 1–5. IEEE.

[Zabell, 1989] Zabell, S. L. (1989). The rule of succession. *Erkenntnis. Kluwer Academic Publishers.*, 31(2-3):283–321.

[Zheng et al., 2009a] Zheng, Y., Chen, Y., Xie, X., and Ma, W.-Y. (2009a). Geolife2.0: A location-based social networking service. *yuzheng, v-yukche, xingx, wyma-@microsoft.com.*

[Zheng et al., 2011] Zheng, Y., Zhang, L., Ma, Z., Xie, X., and Ma, W.-Y. (2011). Recommending friends and locations based on individual location history. *ACM Trans. Web*, 5(1):5:1–5:44.

[Zheng et al., 2009b] Zheng, Y., Zhang, L., Xie, X., and Ma, W. Y. (2009b). Mining interesting locations and travel sequences from GPS trajectories. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 791–800, New York, NY, USA. ACM.

[Zonoozi and Dassanayake, 2006] Zonoozi, M. M. and Dassanayake, P. (2006). User mobility modeling and characterization of mobility patterns. *IEEE J.Sel. A. Commun.*, 15(7):1239–1252.

List of Tables

List of Figures

Danke

Mein besonderer Dank gilt all denen, die mich in irgendeiner Weise während der Promotionsphase unterstützt haben. In ganz besonderer und herzlicher Weise gilt mein Dank meinem Doktorvater Herrn Prof.Dr.rer.nat. Johann Schlichter, von dem ich fachlich und zwischenmenschlich sehr viel lernen durfte, sowohl während meines Informatikstudiums, als auch während meiner Dissertation. Ich danke ihm auch aufs trefflichste für das Interesse, dass er unserer Familie entgegengebracht hat.

Einen ganz besonderen Dank gilt Frau Prof. PhD. Gudrun Klinker für Ihr Gutachten meiner Dissertation, sowie für das Wissen, das ich von Ihr während meines Informatikstudiums erlangt habe.

Ein ebenfalls besonderer Dank gilt dem Herrn Prof. PhD. Bernd Brügge einer Seits für den Vorsitz der Prüfungskommion, anderer Seits ebenfalls für das Wissen, das ich von ihm während meines Informatikstudiums erlangt habe.

In ganz besonderer und herzlicher Weise gilt mein Dank dem Herrn Priv.-Doz. Dr. Georg Groh für sein Gutachten und für das fachliche, sowie zwischenmenschliche Wissen, das ich von ihm erlangt habe. Seine Unterstützung hat massgeblichen Anteil daran, dass meine Dissertation in dieser Form entstanden ist. Außerdem möchte ich ihm für seine zwischenmenschliche Unterstützung, Interesse und Teilnahme an unserem Leben danken, in dem Wissen, dass das nicht gebührend gedankt werden kann. Ich schätze mich sehr glücklich so einen Menschen kennengelernt zu haben.

Meine liebe Frau Lana Bapierre und mein Sohn Leon Hoger Bapierre waren meine größten Unterstützer, Antriebs- und Motivationsgeber. Ich danke meiner lieben Frau Lana Bapierre aus tiefstem Herzen für Ihre Geduld, Zuneigung, Liebe, sowie mentale und emotionale Unterstützung, denn sie hat sehr viel Verantwortung übernommen, damit ich mehr Zeit für meine Dissertation habe. Meine Frau und mein Sohn sind das beste, was mir im Leben passiert ist.