A Probabilistic Approach to Socio-Geographic Reality Mining

THÈSE N^{O} 5018 (2011)

PRÉSENTÉE LE 20 MAI 2011

À LA FACULTÉ SCIENCES ET TECHNIQUES DE L'INGÉNIEUR

LABORATOIRE DE L'IDIAP

PROGRAMME DOCTORAL EN INFORMATIQUE, COMMUNICATIONS ET INFORMATION

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Katayoun FARRAHI

acceptée sur proposition du jury:

Prof. J. D. R. Millan Ruiz, président du jury
Dr D. Gatica-Perez, directeur de thèse
Prof. P. Frossard, rapporteur
Dr J. Laurila, rapporteur
Prof. A. Pentland, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2011

Résumé

Notre façon de vivre au quotidien transmet des indices sur notre vie à notre environnement. Des personnes, mais aussi des appareils électroniques nous entourent en permanence et peuvent percevoir des éléments de notre vie quotidienne. Nos téléphones portables, par exemple, détectent de façon continue nos mouvements et interactions. De telles données sociogéographiques pourraient ainsi être enregistrées en continu par des centaines de millions de personnes à travers le monde et leur exploitation s'avère prometteuse pour révéler des éléments du comportement humain de manière inédite. L'impact sociétal de l'analyse de données comportementales à grande échelle à partir de données d'appareils mobiles est potentiellement important. Comprendre les mouvements et interactions d'une communauté pourrait par exemple permettre d'adopter des mesures appropriées pour prévenir la menace d'une épidémie. L'étude de bases de données de grande envergure centrées sur l'être humain requiert des modèles et outils mathématiques de pointe. Dans cette thèse, nous explorons des modèles probabilistes appelés « topic models » et des méthodes d'apprentissage automatique pour l'analyse d'activités sociogéographiques à grande échelle.

Nous commençons par nous intéresser à deux types de « topic models » pour l'exploitation de données de localisation obtenues à partir de téléphones portables. Nous proposons une méthode basée sur le principe de LDA (« Latent Dirichlet Allocation »), puis sur un modèle appelé « Author Topic Model » afin de découvrir des routines dominantes de localisation, en utilisant la base de données « MIT Reality Mining » qui répertorie les activités de 97 individus sur une période de 16 mois. Nous présentons les nombreuses possibilités qu'offre notre approche pour modéliser les activités. Grâce à notre approche, il est par exemple possible de différencier les utilisateurs ayant des styles de vie très ou peu variés, respectivement, ou de déterminer quand les activités d'un utilisateur fluctuent en fonction du temps par rapport à son activité normale.

Nous considérons ensuite à la fois la localisation et les interactions en ayant recours à des informations mono- et multimodales provenant de connexions Bluetooth ou à des antennes relais afin de découvrir des activités de routine comprenant des informations sur les interactions journalières et les lieux visités. Nous proposons également une méthode de prédiction de données multimodales manquantes basée sur le principe de LDA. Nous considérons par la suite une approche supervisée pour déterminer le type de jour ou d'étudiant en utilisant des données sociogéographiques similaires.

Dans la suite du manuscrit, nous introduisons deux nouvelles approches probabilistes pour corriger certaines limitations de l'approche LDA pour la modélisation d'activité. L'approche précédente souffre de la croissance exponentielle des données et de l'espace des paramètres

lorsqu'on l'applique à des activités de durées importantes ou variables. Nous proposons tout d'abord un « Multi-Level Topic Model » qui permet de modéliser les activités de durée variable. Le « Pairwise-Distance Topic Model » que nous introduisons par la suite permet de capturer des topics représentant des activités de longues durées.

Finalement, nous présentons l'application de notre approche pour l'étude des facteurs influençant les changements d'opinion politique chez les individus. Nous utilisons la base de données « Social Evolution Project Reality Mining » et étudions aussi l'utilité d'autres données comme le journal d'appel des personnes. Nous nous intéressons aux différences de comportement entre les personnes qui changent d'avis et celles qui ne changent pas. Nous combinons différents types de données provenant de différentes modalités pour former une caractéristique « d'exposition » représentant le degré d'exposition d'une personne aux opinions politiques des autres. Nous utilisons les méthodes introduites dans le manuscrit pour étudier le comportement des personnes et des groupes face à l'exposition aux opinions des autres. De cette étude nous dérivons un ensemble de caractéristiques qui permettent de distinguer, de manière statistiquement significative, les personnes qui changent d'opinions et celle qui ne changent pas. Nous étudions aussi dans quelle mesure la différence « d'exposition » à l'opinion des autres influe sur l'intérêt qu'a une personne pour la politique.

En conclusion, cette thèse aborde plusieurs questions importantes du domaine récent de l'informatique pour les sciences sociales (« Computational Social Science »). Nous utilisons de méthodes rigoureuses, fondées sur des principes mathématiques et traitons de données variées provenant de téléphones mobiles. Ceci permet l'extraction d'activités humaines dans des scénarios réels à grande échelle.

Mots Clés : modélisation d'activité, donnée provenant de téléphones mobiles, informatique pour les sciences sociales, apprentissage automatique

Abstract

As we live our daily lives, our surroundings know about it. Our surroundings consist of people, but also our electronic devices. Our mobile phones, for example, continuously sense our movements and interactions. This socio-geographic data could be continuously captured by hundreds of millions of people around the world and promises to reveal important behavioral clues about humans in a manner never before possible. Mining patterns of human behavior from large-scale mobile phone data has deep potential impact on society. For example, by understanding a community's movements and interactions, appropriate measures may be put in place to prevent the threat of an epidemic. The study of such human-centric massive datasets requires advanced mathematical models and tools. In this thesis, we investigate probabilistic topic models as unsupervised machine learning tools for large-scale socio-geographic activity mining.

We first investigate two types of probabilistic topic models for large-scale location-driven phone data mining. We propose a methodology based on Latent Dirichlet Allocation, followed by the Author Topic Model, for the discovery of dominant location routines mined from the MIT Reality Mining data set containing the activities of 97 individuals over the course of a 16 month period. We investigate the many possibilities of our proposed approach in terms of activity modeling, including differentiating users with high and low varying lifestyles and determining when a user's activities fluctuate from the norm over time.

We then consider both location and interaction features from cell tower connections and Bluetooth, in single and multimodal forms for routine discovery, where the daily routines discovered contain information about the interactions of the day in addition to the locations visited. We also propose a method for the prediction of missing multimodal data based on Latent Dirichlet Allocation. We further consider a supervised approach for day type and student type classification using similar socio-geographic features.

We then propose two new probabilistic approaches to alleviate some of the limitations of Latent Dirichlet Allocation for activity modeling. Large duration activities and varying time duration activities can not be modeled with the initially proposed methods due to problems with input and model parameter size explosion. We first propose a Multi-Level Topic Model as a method to incorporate multiple time duration sequences into a probabilistic generative topic model. We then propose the Pairwise-Distance Topic Model as an approach to address the problem of modeling long duration activities with topics.

Finally, we consider an application of our work to the study of influencing factors in human opinion change with mobile sensor data. We consider the Social Evolution Project Reality Mining dataset, and investigate other mobile phone sensor features including communication logs. We consider the difference in behaviors of individuals who change political opinion and those who do not. We combine several types of data to form multimodal exposure features, which express the exposure of individuals to others' political opinions. We use the previously defined methodology based on Latent Dirichlet Allocation to define each group's behaviors in terms of their exposure to opinions, and determine statistically significant features which differentiate those who change opinions and those who do not. We also consider the difference in exposure features of individuals that increases their interest in politics versus those who do not.

Overall, this thesis addresses several important issues in the recent body of work called Computational Social Science. Investigations principled on mathematical models and multiple types of mobile phone sensor data are performed to mine real life human activities in large-scale scenarios.

Keywords: probabilistic topic models, latent dirichlet allocation, Reality Mining, activity modeling, mobile phone data, computational social science, machine learning.

Contents

List of Figures

List of Tables

Acknowledgements

There are many people to which I am indebted during this last step in my seemingly life-long journey as a student!

Foremost, I would like to thank my advisor Daniel. I have always appreciated his kindness, support, guidance, and friendship over the last four years and could really not have asked for a better PhD advisor.

Thank you to Prof. Pentland, Prof. Frossard, Dr. Laurila, and Prof. Millan for kindly agreeing to be on my defense committee. I appreciate your time and comments for the thesis. Thank you to Idiap Research Institute for providing great facilities and opportunities. Also thanks to MULTI for supporting my research and to NRC (Nokia) Lausanne for collaboration opportunities.

So many friends came and then moved on from Idiap during my time here. Hayley, Sileye, Tamara, Stephanie, Elisa, Eileen, Ricardo, I hope we keep in touch! I'm already missing our french lunches; thanks to the professor Stephanie, and the fellow pupils Hayley, Elisa, and Nik. Thanks to all the Idiap girls for our lunches. Thanks to the Iranian community for helping me practice my farsi, notably Gelareh, Samira, Hamed, Majid, Afsaneh, and the list goes on. CC and Alex, I will miss our gossiping sessions. Thanks to my lunchtime badminton buddies and to Friday night billiards players as well!

Big thank you to Iacopo and Corinne who helped us in many ways and showed us some special places in Switzerland. Special thanks to Kevin and Letizia for their friendship and for hours of board game fun, especially during 'on-call' days in Bern. Thanks to Daniel and Jennifer for their friendship.

It was a real pleasure working with all of Daniel's group members. Dinesh, Hayley, Joan, Paco, Oya, Hari, Radu, Dayra, Minh-Tri, 'Apple' Gokul, Bogdan, Laurent. Thanks for many interesting discussions over group lunches and group activities; I will definitely miss the group! I also learned many things from Jean-Marc and his group, especially Remi, CC, Alex, Samira, Jagan. Besides work, you taught me my hiking capabilities! Thanks to Minh-Tri, Remi, and Sileye for many insightful work discussions. You've helped me alot in my thesis and are my mentors.

I have always enjoyed the atmosphere in our office thanks to Sileye, Hayley, Dinesh, Radu, Chris, Stephanie, Stefan, Venky, Majid, Alex, CC, Remi, in chronological order. Special 'thanks' (in quotes) to Radu for his office shenanigans.

It was a great opportunity for me to visit the HD lab at MIT as an intern. Thanks Anmol and Sandy for hosting me. I learned alot during my short stay there, and met many inspirational people, including Sandy's students Anmol, Wei, Manuel, Taemie, Ben, Daniel, Riley, Wen. Shirin and Arya, I will never forget your hospitality during my stay at your place for so long, thank you!

Finally, a big thank you to all my dear family and close family friends for their love and unconditional support. My dear aunts and uncles, my amazing grandma, and my dear cousins, too many to name here, love you all! Thanks Berdjis Khanoom and Shirin jan for your close friendship. Thanks to my 'lil sis Shirini and Arya. Love you very much! Thanks to Faafaa and Foofoo :), Laura-Shirin for being so cute, Claudia, Nima, Parvin Khanoom and Aghaye Doctor for their support. Mom and dad, I can't express in words how thankful I am to you both. Madar I am sooooo proud of you, and Daddang pedat sham.

Last but not least, thank you Farzad for everything! You are really one of a kind, thanks for your love and support! I will cherish it always.

Chapter 1

Introduction

Sensors are everywhere, continuously gathering information as we live our daily lives. We live in a technology driven society where, Lazer et al [57] stated it best, each one of us continuously leaves digital traces behind. Whether using email, the telephone, a bank machine, or even simpler activities such as driving, using a photocopy machine, and a camera, all of these activities leave traces of our behavior. Recently, these devices have been viewed from an engineering perspective as sensors, capturing data which scientists in many disciplines are very excited about. This data potentially impacts everyone of us as researchers begin to study the possibilities of their use. Applications to society as a whole are being investigated in terms of epidemiology and psychology [64], urban planning [37], security, and even in the analysis of poverty [95].

Reality Mining, pioneered by Nathan Eagle and Alex Pentland, is the study of human social behavior based on wireless mobile phone sensed data. Mobile phones are particularly promising as sensors due to their vast usage over the world on a daily continuous basis, and also due to the numerous types of sensors embedded in the device. Not only do people carry them around as they live their daily lives sensing their location, and motion (via the accelerometer), their interactions can also be captured by Bluetooth, not only with other individuals carrying phones, but also with computers in proximity. There are many other forms of human behaviors that can be sensed with mobile phones, one of which is the reason they were invented in the first place, for communication. The mobile phone has developed, due to its paramount nature, from a simple communication device to include many other tools such as a cameras, browsers, games, calendars, alarm clocks, and will surely continue to develop in the future. All of these forms of data can be analyzed to reveal details about human behavior.

The focus in this thesis is on large-scale socio-geographic data obtained by mobile phone sensors capturing real-life location and proximity data. Our goal is to mine human activities and routines from this socio-geographic data. We define routines to be temporal regularities in peoples' lives. Activity modeling is a relatively recent domain in computer science, and of great interest in a new discipline named Computational Social Science [57].

1.1 Current Challenges

As large scale data collections on human behavior become more readily available, the need for effective methods and mathematical models for analysis becomes crucial in order to make good use of the sources. In machine learning, algorithms have been developed to recognize complex patterns and make intelligent decisions based on data. Traditional machine learning models are recognized as useful tools for large-scale data analysis [5, 22]. They have been used in the domain of human behavior analysis, though their limitations with new types of data and human-centric questions become apparent. For example, many of the traditional machine learning models are supervised, requiring training data which is often impossible or illegal to collect on human subjects. Other specifications related to human-centric data include the multimodal aspect, the noise, the massive quantity, and the complex questions of interest.

More specifically, data collected by mobile phone sensors include many types, ranging from GPS, Bluetooth, accelerometer, to voice features. Each of these sensors may be sampled with varying frequencies, each has varying timescales and differing characteristics, and each has its own sources of noise. For example, a person can interact with more than one other person at a given time; some potential features to model include the group size, identity, and relationship with those in proximity. Noise is inherent in human behavior in relation to sensors. People forget their phones, lend them to friends, set the time incorrectly, forget to turn on Bluetooth, and most importantly the phone is not always attached to the individual [76]. Each sensor has its particular problem. For example GPS does not work indoors and Bluetooth detects devices through thin walls. Continuous data on large populations pose many computational challenges. For example traditional techniques using linear algebra are not easily applicable due to large matrix operations. Finally, and most importantly, traditional machine learning techniques have not been designed to target the questions of interest. For example, a typical question of interest would be "what are the differences in the daily routines between two populations (e.g., of students)?". We do not know what machine learning learning tools would best solve this problem. There are several state-of-the art classification tools to divide the groups with lowest error, however, we are interested in finding the similar and differentiating features within these groups and understanding how significant they are.

We believe machine learning methods can provide useful algorithms for large-scale activity modeling, as we show in this thesis. However they do present their limitations in face of human-centric data and can be built upon. Though machine learning techniques are very useful techniques for large-scale human behavior analysis, we believe they pose limitations and can be used as a basis for new methods targeting the field of Computational Social Science.

In this thesis we choose to investigate probabilistic topic models as the basic tool. Topic models are chosen first and foremost for their unsupervised nature. They have several other advantages making them an attractive choice for the goals of this thesis. Their probabilistic, generative nature make them attractive over discriminative approaches since we are interested in modeling the nature of the data. Further, the basic statistical structure of the model is intuitive and provides opportunity for extensions with Markov Chain Monte Carlo (MCMC) methods for inference. The input feature structure in topic models, called bag of words, is critical in removing redundant time information and is key in handling

the large amounts of noise in the data. Very importantly, topic models can be applied to very large data collections. They also can be manipulated in various ways to integrate multiple data types.

1.2 Summary of Contributions

The main contributions of this thesis are summarized below.

(1) Activity modeling with topic models, an unsupervised approach. We propose a fully unsupervised approach for human activity discovery based on two existing probabilistic topic models, Latent Dirichlet Allocation [10] and the Author Topic Model [82], which we published in [31]. We also present an approach to both individually and jointly model a user's locations and interactions in a manner suitable for robust human activity mining and present a method to predict missing socio-geographic data, essentially validating our topic model approach to jointly model the multimodal data. These ideas were published in [30, 32]. We thoroughly analyze the activity recognition tasks that can be achieved with our proposed method. The tasks include determining dominant routines on certain days, finding a user's dominant routines, differentiating between users that live high entropy versus low entropy lives, determining when a large change occurs in a user's typical routine over time, and discovering groups of users that follow certain trends. All of these activity modeling experiments were performed on the Reality Mining data [26] and detailed journal versions are published in [34, 35].

(2) Supervised approach to evaluate feature selection validity. We use a supervised approach, namely Support Vector Machines, to validate the feature selection in terms of both location and interaction data addressing two tasks with prior knowledge of labels: day types (weekend versus weekday) and student types (engineering versus business). This was published in [29].

(3) Proposal of the Multi-level Topic Model. We address a limitation of probabilistic topic models, namely incorporating varying time durations without vocabulary size explosion, by proposing a new extension to Latent Dirichlet Allocation. We call the new technique the Multi-Level Topic Model. It is formally defined, implemented, and analyzed experimentally. Some of these ideas were published in [33].

(4) Proposal of the Pairwise-Distance Topic Model. We address a similar problem, this time focusing on the modeling of very large length sequences (or large n-grams), with a new graphical model inspired by Latent Dirichlet Allocation which we call the Pairwise-Distance Topic Model. We present the inference procedure details, as well as the results of the implementation. This work is new and has not yet been published.

(5) Political Opinion Change Modeling. Finally, we address an important question in the Social Sciences, that of influencing behavior in opinion change. Using our proposed methods for activity mining with Latent Dirichlet Allocation, we address a very multimodal data collection to incorporate interaction, communication, relationship, as well as opinion features to determine the difference in

the routines of people who change their political opinion over a three month period versus those who do not. This was published in [66].

1.3 Thesis Outline and Problems Addressed

We begin the thesis by discussing the related work in Chapter 2. The previous works in activity recognition are presented, grouped by the various possible types of sensors. We begin with indoor sensors, including computer vision problems, followed by wearable devices. We present a separate section on related works using mobile phone sensors in the section titled Reality Mining, also a wearable device. Lastly we present topic models and their development as well as related papers in people-centric networks (social networks) as well as n-gram models.

In Chapter 3 we present an overview of our methodology based on topic models for activity modeling. The goal of this chapter is to investigate the role of the topic model as a tool for location-driven activity modeling and to present the possibilities of use. We describe the components in detail, beginning with the bag representation, followed by the Latent Dirichlet Allocation and the Author Topic Model. The experimental results are presented for large-scale location data in Section 3.4, as well as the details of various types of tasks that can be performed with the methodology. Perplexity is used for model parameter selection, as detailed in Section 3.4.2. Resulting topics for location activities are differentiated from the traditional clustering k-means model in Section 3.4.6.

A multimodal data representation for the joint modeling of socio-geographic data is presented in Chapter 4. This chapter addresses the problems of handling multimodal data in terms of both discovery, prediction, and classification with supervision. The framework for modeling this multimodal data is presented, in addition to a framework for the prediction of missing data. Experiments and results follow including analysis for predicting missing socio-geographic data in Section 4.3.4. In Section 4.4 we evaluate the individual location and proximity features as well as multimodal representations of these features with the supervised discriminative technique of support vector machines. We show classification results on two differing tasks, the day type and the student type to show the features in question are discriminative.

We present a limitation of the topic modeling technique for activity recognition and propose two new models to overcome these limitations in Chapter 5. In Section 5.2 we present the Multi-Level Topic Model as a way to overcome the problem of modeling varying time duration activities simultaneously without vocabulary size explosion. This model is formalized, then experiments are presented including comparisons of sequences discovered with the most frequently occurring sequences in the data. In Section 5.3 we present the Pairwise-Distance Topic Model, a probabilistic generative approach inspired by Latent Dirichlet Allocation to handle the problem of large duration activity discovery. The details of the inference procedure are presented followed by the experimental results and analysis.

Finally, in Chapter 6 we apply the general activity modeling techniques proposed in this thesis to address an important question in the social sciences, that of opinion change. We investigate, in collaboration with Anmol Madan and Alex Pentland at the MIT Human Dynamics Lab, the differences in activities of

individuals who change their political opinions versus those who do not. We use multimodal features containing communication, interaction, relationships, as well as opinions with an approach based on Latent Dirichlet Allocation to study the differences and similarities between these two groups of individuals.

Chapter 2

Related Work

2.1 Overview

This chapter reviews the most relevant previous work related to this thesis. First in Section 2.2 we present the prior art in the field of activity recognition. Activity recognition, also referred to as activity modeling, is a relatively new domain applicable to various types of data including indoor sensors and wearable devices. In Section 2.3, we focus on previous works conducted on large-scale mobile phone sensed data, termed Reality Mining. This prior work generally aims at understanding human and social behavior on a large-scale which was previously not possible with other sensor data types due to scale and availability. The machine learning methodology we apply and build upon in this thesis for activity discovery (what we refer to as socio-geographic Reality Mining data), is the family of probabilistic models known as topic models. In Section 2.4 we review the key works in this domain and provide a brief history of their development. We also discuss extensions of these models which were motivating ideas for this thesis.

2.2 Activity Recognition

Even if a system cannot fully model a user's beliefs, desires, and intentions, it can still be useful if it can recognize some of his/her activities [17]. Choudhury et al. summarize the techniques and challenges of activity modeling in [17], which serve as a good introduction to this area of research. An activity recognition system has three main sub-components also shown in Figure 2.1:

1. A low-level sensing module that gathers relevant information about activities, e.g., camera, microphone, or motion sensors. In Section 2.2.1 and 2.2.2 we discuss the types of activity recognition tasks that can be addressed with the various categories of sensing devices, not including mobile phones. In Section 2.3 we focus on previous works addressed with mobile phones as sensors.

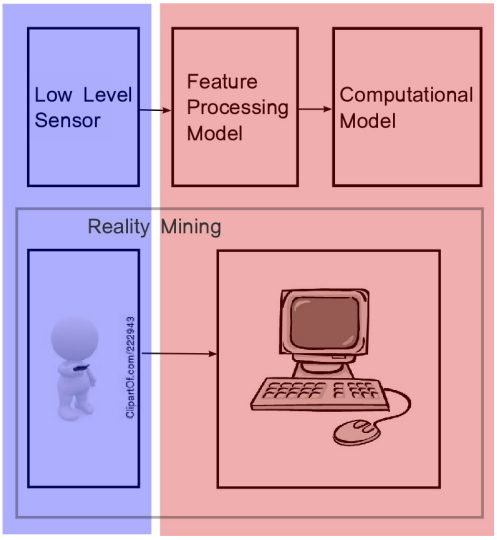


Figure 2.1. Overview of an activity recognition system. The three main sub-components are the low-level sensing module, the feature processing and selection module, and the computational model that infers user activities. Reality Mining systems use the multiple sensors on the mobile phones to capture data. The rest of the processing is partly done on a computer and partly on the device.

2. A feature processing and feature selection module that processes the raw sensor data into features that can help discriminate between activities.
3. A computational model that uses these various features to infer the activity that an individual or group is engaged in.

In the following sections, we discuss some of the computational models that combine steps 2 and 3, or only step 3, to transform the rich and complex set of raw sensor data into higher-level descriptions of people and/or group activities and interactions. Section 2.4 focuses particularly on one generation of promising computational models in machine learning, topic models.

A critical component of an activity modeling system is the sensing module that gathers relevant information about activities. There are various categories of sensors other than mobile phones (which is the focus of the next section) including indoor fixed sensors such as cameras, microphones, motion sensors, and wearable computing devices. Each category of sensing devices has advantages and disadvantages and can be employed for specific activity recognition and discovery tasks. Multi-modal data can be collected to obtain a wider range of information (e.g., audio + video), sometimes complicating the activity recognition task, however, enriching the information available for computation.

2.2.1 Indoor Sensors

Indoor sensors include cameras, microphones and proximity or motion sensors. These are the most commonly used sensors, with many mature processing techniques related to feature extraction for activity characterization. The restrictions with these types of sensors are that they are often fixed and only those

activities that occur in the local physical space covered by the sensors can be recognized. In the next paragraphs, we give examples of existing work using fixed indoor sensor data.

There is a great deal of research on human activity recognition and modeling in the area of computer vision. We mention a few works for the sake of completion. Topic models have been used in vision research for activity recognition from video data [72, 94]. In [72], human action categories such as "walking", "jogging", or "boxing" are discovered with Probabilistic Latent Semantic Analysis [45] in which topics correspond to action categories. In [94], activities, are discovered in complex video scenes, such as crowded traffic scenes, with many activities occurring simultaneously. The methodology built from topic models is also used to discover unusual activities and interactions, an example of which is "a pedestrian crossing the road outside of the sidewalk". Another unsupervised methodology for activity discovery is [41], where suffix trees are used for activity discovery.

The indoor spaces in the daily life of an individual typically begin with their household. Household activities have been automatically discovered on a large scale using event-streams and an undirected graph approach, considering all the rooms in a house [40], as well as on a small scale using a switching Hidden Semi-Markov Model, recognizing kitchen activities (such as making coffee, eating breakfast) [24]. PlaceLab [55] is an example of a "living lab", where hundreds of sensors are built into objects and the home environment for various research purposes including activity recognition [49, 61, 86].

The second most common indoor space an individual frequents is the office. The literature reviewed for activity recognition in office spaces mostly used a Markov model approach to recognize activities, for instance, on a multi-room scale [96], or in a smaller meeting room scale [68]. The work by Wren et al [97], models office activity as a Markov process, to detect deviations of short-term behavior patterns from long-term patterns. The method finds time periods where the dynamics of the group behavior deviate significantly from "usual" behavior, successfully identifying holidays, periods of vacation, and periods of organizational disruption to the group (e.g., senior manager change). The time period of the dataset is comparable to the one we are investigating, covering the course of a year. There is also on the order of one hundred people in the study, though their identity is completely unknown due to the sensing strategy.

2.2.2 Wearable Devices

There is an increasing body of work on activity recognition using various types of wearable sensors. A pioneering work in wearable sensing was done by Choudhury and Pentland [14] using wearable database acquisition boards called the Sociometer. In [14], the audio information was stored from a microphone for twenty three people and about 66 hours each over a two week period. The data was recorded both indoors and outdoors. The sociometer is used to capture face-to-face interactions and the data is measured in order to learn the structure and dynamics of human communication networks. The device contains an accelerometer, a microphone, and an IR sensor. In [15] Choudhury and Basu use data collected by the sociometer to quantitatively investigate the ways in which a given person influences the joint turn-taking in a conversation using a Mixed-Memory Markov Model. In [16], methods to automatically learn social network structures are developed using sociometer data. A larger data campaign capturing spontaneous

face-to-face conversation was done by Wyatt et al in [98]. Twenty four subjects were given personal digital assistants (PDAs) containing 8 sensors to capture unconstrained and natural conversational features over a 9 month period, resulting in 4,400 hours of data.

Olguin et al [73, 74] also developed wearable electronic badges to measure the amount of face-to-face interaction, conversational time, physical proximity to other people, and physical activity levels. Their sociometric badges have many capabilities, including recognizing human activities such as sitting, standing, and walking, in real-time. In [73], the badges are deployed in a real organization with 22 employees participating over a one month period. The goal is to be able to understand how patterns of behavior shape the individual and the organization. The developed methods are able to reasonably predict employees' self-assessments of job satisfaction with the data sensed, in combination with survey and email information. In [74], the sociometric badges are deployed in a medical facility for health management. In this study, 67 nurses working in the Post-anesthesia Care Unit (PACU) in the Boston area were given badges over a 27 day period, resulting in 3906 hours of data. The authors show that it is possible to identify individual personality traits and measure group performance in this hospital unit with good accuracy.

Wearable sensors have to potential of providing more detailed information and insight to human health related applications. Acceleration and heart rate information have been gathered in [87] for the automatic recognition of physical activities as well as their intensities. Munguia et al [87] use five accelerometers and a wireless heart rate monitor and evaluate their algorithm on 30 physical gymnasium activities collected from 21 subjects. A real-time algorithm has been presented for automatic human activity recognition and in some cases, for detection of the activity intensities. In [54] wearable sensors are used for sleep posture detection, which may be valuable for people with sleep apnea disorders. The experiment concludes that sufficiently accurate estimation of basic sleep postures can be done by a single power-efficient wrist-worn sensor. Work by Madan el al [64] addresses a larger-scale concern in health care, termed behavioral epidemiology, with the goal of monitoring how individual behavior is affected by illness and stress. The wearable sensors used for this study are mobile phones, involving daily surveys. The method successfully predicts the health status of subjects with mobile phone features.

A closely related work published simultaneously to ours [30, 31], is described in [48] and uses topic models for human activity discovery using wearable sensor data. The method identifies activity patterns in one single person's daily life over sixteen days, using two wearable sensors, one placed on the right hip and the other on the right wrist. For activity recognition, the Latent Dirichlet Allocation (LDA) topic model is used where activity words are manually labeled or automatically recognized low-level activities, and the topic model is used to discover patterns in these activities, which are essentially co-occurring low-level activities. In contrast, our work investigates the human routine discovery task from mobile phone data, on a true large scale, and we further use this data to discover group routines in addition to individual routines. Further, our methodology has proven successful on lower level input data which can be obtained more directly from sensor data, such as the locations of an individual and their proximate interactions [30]. The methodology proposed in [48] requires higher level information regarding a person's

activities through the use of multiple devices attached to various body parts, unlike our work which only relies on one single device (a phone) which is worn and used naturally.

2.3 Reality Mining

The mobile phone is a very unique sensing device continuously capturing our location, interaction, communication, and motion traces left behind in our daily lives [57]. Mobile phone technology allows us to capture data related to the daily routines of large numbers of people over very long periods. More specifically, their locations, such as being at work or home, can be learned and recognized from location sensors [26, 52, 44]. Interactions can be captured by Bluetooth, which detects other Bluetooth devices within a small radius [26, 52, 63, 66]. Phone call and SMS activities can further be recorded [26, 69, 52, 63, 66, 95]. Phone application usage can be saved including the camera, calendar, games, and web browser usage [69, 52]. Finally, content, including photos and video, can also be collected [20, 69]. Researchers are just beginning to understand the implications of such data collections for fields ranging from epidemiology [91] and dynamical network analysis [43] to understanding the dynamics of slums [95].

Research using mobile phone data has mostly focused on location-driven data analysis, more specifically, using Global Positioning System (GPS) data to predict transportation modes [77, 80], to predict user destinations [53] or paths [1], and to predict daily step count [84]. Other location-driven tasks have made use of Global System for Mobile Communications (GSM) data for indoor localization [75] or WiFi for large-scale localization [58]. There are several works related to activity modeling from location-driven phone sensor data. CitySense [60] is a mobile application which uses GPS and WiFi data to summarize "hotspots" of activity a city, which can then be used to make recommendations to people regarding, for example, preferred restaurants and nightclubs [50]. Liao et al [59] use GPS data traces to label and extract a person's activities and significant places. Their method is based on Relational Markov Networks. The BeaconPrint algorithm [44] uses both WiFi and GSM to learn the places a user goes and detect if the user returns to these places.

Recently, mobile phones have been programmed to capture non-linguistic speech attributes [62, 63]. These non verbal speech features have been used for sound classification (for example music versus voice) and for the discovery of sound events [62]. The VibeFone application[63], uses location, proximity, and tone of voice features to infer specific aspects of peoples' social lives. The mobile application has two special modes, the Jerk-o-Meter and the Wingman3G, in which VibeFone evaluates the subject's speech and provides feedback to subjects. Experiments have been conducted on several small scale data collections to measure and predict interest in conversation, and to measure attraction in a speed-dating scenario.

In [91], a study of how mobile phone viruses spread investigated joint location and proximity mobile phone data. Wang at al [91] model the mobility of mobile phone users to study the spreading patterns characterizing a mobile virus outbreak. They find that Bluetooth viruses spread slowly due to human mobility; however, viruses utilizing multimedia messaging services could infect all users in hours. Other

related works include [27], CenceMe [69], and [95]. In [27] individual calling diversity is used to explain the economic development of cities. Eagle et al find that the diversity of individuals relationships is strongly correlated with the economic development of communities [27]. CenceMe [69] is a personal sensing system that enables activity sharing sensed automatically by mobile phones in a user's online social network. The sensed activities, referred to as "sensing presence", captures a users' status in terms of activities (e.g., sitting, walking), disposition (e.g., happy, sad), habits (e.g., at the gym, at work), and surroundings (e.g., noisy). These features can then be shared in popular social networking sites such as Facebook, Myspace, as well as instant messaging tools such as Skype and Pidgin. Wesolowski and Eagle [95] use mobile call logs collected over a one year period to better understand one of the largest slums, Kibera located in Nairobi, Kenya.

Eagle and Pentland [25], the pioneers in the Reality Mining research domain, used Principle Component Analysis (PCA) to identify the main components structuring daily human behavior. The main components of human activities, which are the top eigenvectors of the PCA decomposition are termed *eigenbehaviors*. To define the daily life of an individual in terms of eigenbehaviors, the top eigenbehaviors will show the main routines in the life of a group of users, for example, being at home overnight. The role of the remaining eigenbehaviors is to describe the more precise, non-typical behaviors in individuals' or the group's lives. Eigenbehaviors are described over what we consider to be fine-grain locations (30-minute time intervals) and are representative of the entire day's activities as opposed to morning only or evening only. Results are presented for location data only where Bluetooth and raw location data is used in an HMM structure to infer a user's location for a given time. In Chapter 3, we present a methodology based on a bag of location sequences structure. This method is advantageous over [25] in that it contains multiple time considerations, can discover activities over parts of the day, can be used in describing user's and groups of user's behaviors, and can be easily applied to very large datasets.

Another closely related work from mobile sensor data is by Gonzalez et al [37]. Recently, they used mobile phone data to study the trajectories of human mobility patterns, and found that human trajectories show a high degree of temporal and spatial regularity, more specifically, that individual travel patterns can "collapse" into a single spatial probability distribution showing that humans follow simple, reproducible patterns. This study was performed on a large-scale mobile phone dataset from a phone operator of over 100 000 users over a period of six months. In [13], phone call data has been used to study the mean collective behavior of humans at large scale, focusing on the occurrence of anomalous events. The authors also investigate patterns of calling activity at the individual level and model the individual calling patterns (time between phone calls) as heavy tailed.

In Table 2.1, we present a summary of several data collections already mentioned in previous paragraphs. The data collection used (and not collected) by Gonzalez et al [37], which is not publicly available, is on the largest scale as it comes from a phone operator, with the drawback of containing location information only when phone communication is present. CenceMe [69] is unique in focusing on the social networking aspect but this work was not intended to record large-scale data. VibeFones [63] and [52] collect audio features from the mobile microphone enabling more personal human studies. The Nokia-Idiap

Table 2.1. Reality Mining

	Eagle and Pentland [26]	Madan et al [66, 64]	Kiukkonen et al [52]	Gonzalez et al [37]	CenceMe [69]	VibeFones [63]
location via cell tower	X					X
location via GPS			X		X	
location via call data				X		
location via WLAN		X	X		X	
interaction via Bluetooth	X	X	X		X	X
communication via call, SMS logs	X	X	X	X		X
microphone audio features			X		X	X
accelerometer			X		X	
number of users	100	70	170	100 000	n/a	several
time period of data collection	16 months	3 months	1 year	6 months	n/a	small scale
daily survey capabilities	No	Yes	No	No	n/a	user feedback capabilities

data collection campaign [52] is very recent and targets richer data collection on large-scale and heterogeneous participants consisting of family and friends, involving over 170 participants over a year of time, providing new opportunities for future works. The two data collections by Madan et al [66, 64] occur over a three month period, though these are differing periods, each with on the order of 70 participants.

2.4 Probabilistic Topic Models

Topic models were initially developed to manage large collections of text documents. They were developed as methods for the automatic organization, management, and retrieval of large modern digital library databases [8]. They have been used more recently for other sources of data such as image [70, 79], video [72], genetics [78], wearable sensor data [48], location [31] and multimodal [66, 34] phone data.

Before probabilistic topic models were developed, Latent Semantic Analysis (LSA) [21] (also called Latent Semantic Indexing) was used for large corpus analysis. LSA uses linear algebra techniques to produce a set of concepts from document/term frequency counts. It suffers from a set of limitations, including the matrix dimensions not always being interpretable, which led to the development of Probabilistic Latent Semantic Analysis (PLSA) [45] proposed by Hofmann (also known as PLSI). PLSA is a more principled approach than LSA, based on a statistical foundation and incorporating a hidden layer called topics. A shortcoming of PLSA is that it is not a proper generative model for new documents. This shortcoming was overcome by Latent Dirichlet Allocation (LDA) [10] proposed by Blei et al, which we define in detail in Chapter 3. Basically, LDA defines a Dirichlet prior on the document/topic and

term/topic distributions. There have been many recent extensions of LDA, several of which we introduce in the following paragraphs. However, LDA has been used as a basis for a vast range of research addressing many types of questions, which we will also introduce in the following paragraphs.

Computation of the true values of the LDA model parameters is intractable. There have been several techniques proposed over the years to approximate them. In the original LDA method, variational Bayesian inference is used to approximate the model parameters [10]. A more popular approach based on Markov Chain Monte Carlo methods, which we also use in this thesis, is based on collapsed Gibbs Sampling [38], where the term collapsed simply refers to integrating out the model parameters in deriving the sampling equation. Another approach is based on collapsed variational Bayesian inference [89]. A very interesting paper by Asuncion et al [2] compares all the inference approaches proposed to solve LDA. The various learning algorithms are empirically compared and shown to be very close in performance, differing mainly in the amount of smoothing applied to the estimated parameters.

Topic models for social networks

On the modeling front, Exponential Random Graph Models (ERGMs) and its extensions [81, 99, 39, 47] have been often used to understand interaction networks. Topic models have also been explored for social network analysis. Steyvers et al [85] extended LDA to explicitly model the authorship of documents in the Author-Topic model and apply this to a large corpus of academic papers from CiteSeer. We also use the Author-Topic model in this thesis for individual activity recognition in a large mobile phone network and describe the model in detail in Chapter 3. Erosheva et al [28] define a generative process not only for words, but also for citations to other documents in a corpus, capturing relations between document entities in the corpus. The model is called the mixed membership model and sometimes referred to as LinkLDA. The Link-PLSA-LDA model [71] models the text content of the citing document and the influences on the document. Yano et al [100] develop a probabilistic generative model based on LDA for the generation of blog posts and comments jointly on a blog site. LinkLDA [28] is applied for political blog post and comment modeling by generating a bag of users who respond to a blog post in addition to generating the words in the blog post from its topic mixture. The authors additionally propose the CommentLDA generative model to generate the content of blog post comments.

Another set of related papers developed for social network analysis is the Group-Topic model [92], the Author-Recipient Topic model (ART) [67], and the Role-Author-Recipient Topic model (RART) [67]. The ART model builds on the Author-Topic model for social network analysis, adding the key attribute that the distribution over topics depends on both the sender and recipient, steering the discovery of topics according to the relationships between people. The RART model also includes peoples' roles. Finally, the Group-Topic model clusters entities with relations between them, as well as attributes of these relations.

Other Topic Models

There have been many other extensions of LDA. Among models of great interest in this domain, we can list the Nested Chinese Restaurant Process [11], Dynamic Topic models [6], Supervised Topic

models [9], Correlated Topic models [7], and Hierarchical Dirichlet Processes [88]. Each of these models explores one or more improvements to the basic methodology, to explicitly take into account dynamics, hierarchies, and correlations in data.

Topic models and n-grams

In this thesis we are particularly interested in related works on the topic of n-gram models since we propose two approaches of our own in Chapter 5. In text modeling, term collocation has been tackled extensively, though usually trigrams (a word level trigram contains three consecutive words) are the maximal n-gram considered [4]. The simplest method is based on counting. In [51], word frequency is combined with linguistic knowledge to discover meaningful phrases. In [83], collocation discovery of words is based on variance. Other authors [19, 23] use hypothesis testing to assess whether or not two words occur together more often than by chance. These methods are of particular interest in text analysis, however, none of these methods would be relevant for large n-gram discovery.

Probabilistic topic models have been used for n-gram discovery. The bigram topic model [90], the LDA collocation model [93], and the topical n-gram model [93] are all extensions of LDA to tackle this problem. The bigram topic model [90] generates bigrams where the generative process first draws a topic then words are generated from a distribution dependent on the topic and previous word. The LDA collocation model [93] introduces a new set of random variables into the model that denote if a bigram can be formed with the previous word or not. The topical n-gram model [93] is an extension to the LDA collocation model, and is more general than the bigram model. This approach retains counts of bigram occurrences and thus could not easily be extended for very large n due to matrix size explosion. The main advantage of the topical n-gram model over the LDA collocation model is that it takes context into account when determining whether or not to form a bigram. However, in the LDA collocation model, once a pair of words is considered to form a bigram, it is always a bigram without topic dependency.

2.5 Conclusion

In this chapter we summarized the papers most relevant to this thesis, starting with related works in activity recognition from various types of sensors. We then present probabilistic topic models, including their development and state-of-the-art models on this subject in machine learning.

Chapter 3

Location-Based Activity Recognition

In this chapter, our goal is the discovery and analysis of human routines which characterize both individual and group behaviors in terms of location patterns. We develop an unsupervised methodology based on two probabilistic topic models and apply them to the daily life of 97 mobile phone users over a 16-month period to achieve these goals. As described in Chapter 2, topic models are probabilistic generative models for documents that identify the latent structure that underlies a corpus of words. We define routines to be temporal regularities in people's lives. A routine often involves patterns of location transitions over time (e.g., being at work or going from work to home), possibly over different time scales and for varying time intervals. Routines dominating the entire group's activities, identified with a methodology based on the Latent Dirichlet Allocation topic model, include "going to work late", "going home early", "working non-stop" and "having no reception (phone off)" at different times over varying time intervals. We also detect routines which are characteristic of specific users, with a methodology based on the Author-Topic model. With the routines discovered, and the two methods of characterizing days and users, we can then perform various tasks. We use the routines discovered to determine behavioral patterns of users and groups of users. For example, we can find individuals that display specific daily routines, such as "going to work early" or "turning off the mobile (or having no reception) in the evenings". We are also able to characterize daily patterns by determining the topic structure of days in addition to determining whether certain routines occur dominantly on weekends or weekdays. Furthermore, the routines discovered can be used to rank users or find groups of users who display certain routines. We can also characterize users based on their entropy. We compare our method to one based on clustering using k-means. Finally, we analyze an individual's routines over time to determine regions with high variations, which may correspond to specific events.

Our focus is the automatic discovery of human activities and routines from mobile phone location data collected by one hundred individuals over the course of a year. Automatic routine classification and discovery are in general challenging tasks as people's locations often vary from day to day and from individual to individual, and data from sensors can frequently be incomplete as well as noisy. A supervised learning approach to activity recognition would require prior knowledge in the form of predefined activity

categories and labeled data [59]. In contrast, an unsupervised learning approach has the potential of automatic discovery of emerging routines of people not requiring training data. Through discovery, sifting through large amounts of noisy data becomes possible. Further, one can cluster data (i.e., people or days) corresponding to the most common routines (those of several people) and discover the dataset structure with minimal prior knowledge. We show that topic models prove to be effective in making sense of behavioral patterns at large-scale while filtering out the immense amount of noise in real-life data.

In this chapter we present an overview of our approach to large scale human activity mining from location data. We present details of each component, beginning with the bag construction in Section 3.2. We then present the topic models used with our approach, first focusing on Latent Dirichlet Allocation, followed by the Author Topic Model. Finally we present the experiments and results in Section 3.4.

The work presented here was originally published in modified form in [31] and a longer journal version was published in [35].

3.1 Overview of Our Work

We use the Reality Mining dataset [26], which contains the mobile phone sensor data recordings of 97 subjects studying or working at MIT over the 2004-2005 academic year. As discussed in Chapter 2, the dataset contains the location (cell tower connections), proximity (Bluetooth connections), communication as well as phone application usage of the subjects, though much of this data is noisy and missing. Here we focus on the location dataset, which is given by cell tower connections. Throughout the study over 32 000 towers were recorded, to which we assign semantic labels. We assign 'location labels' of *home* (H), *work* (W), or *other (or out)* (O) to the towers using labels provided by the collectors of the dataset. More precise details regarding location labeling and the dataset can be found in Sections 3.2.1 and 3.4.1, respectively. At this point, we simply assume that we can represent a day in the life of a mobile phone user in terms of location labels for visualization and description purposes. Assuming we can express the day in the life of a person's locations in terms of these labels, with the addition of a fourth *no reception* (N) label in case the phone was off or no data was recorded, we can then visualize the users' location patterns as a function of time of day, as in Figure 3.1 (a) and (b). Each row in the figures is a day of a person's life in terms of his/her location, where the x-axis is the time of day, and the four colors represent the four location labels. Figure 3.1 (a) shows our entire dataset for the 97 users and their 491 days of activities, many of which contain *no reception* the entire day. Figure 3.1 (b) shows the input dataset used in which we remove days containing entirely *no reception* labels. Looking at Figure 3.1 (b), the immense quantity of noise and missing data becomes apparent as well as the amount of data and complex mixture of activities which potentially exist. In addition, it is not apparent how to determine dominating group routines and how to characterize individuals in terms of the groups' routines. These are a few of the points we address with our proposed methodology, illustrated in Figure 3.2.

Our overall goal is to determine what human routines are contained in mobile tower connection data and how to discover them in an unsupervised manner. As described earlier, we represent a day in the life

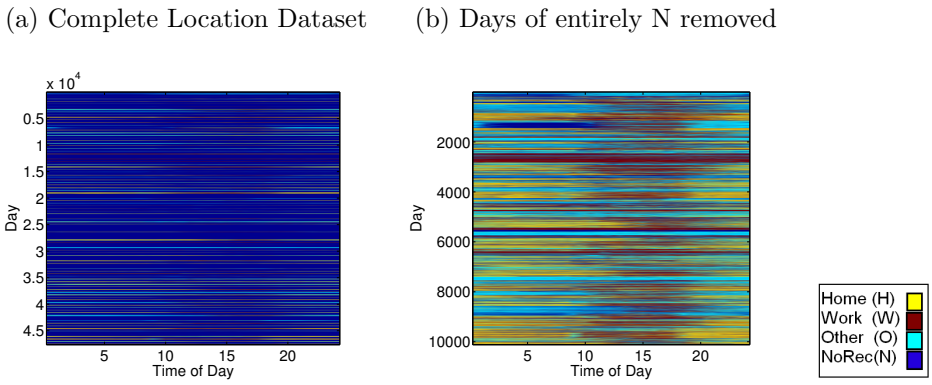


Figure 3.1. Visualizations of the location data for (a) all the users and the entire set of days and (b) all the users and days excluding days which contain entirely no data. The x axis corresponds to the time of day (in hours). The y axis corresponds to days.

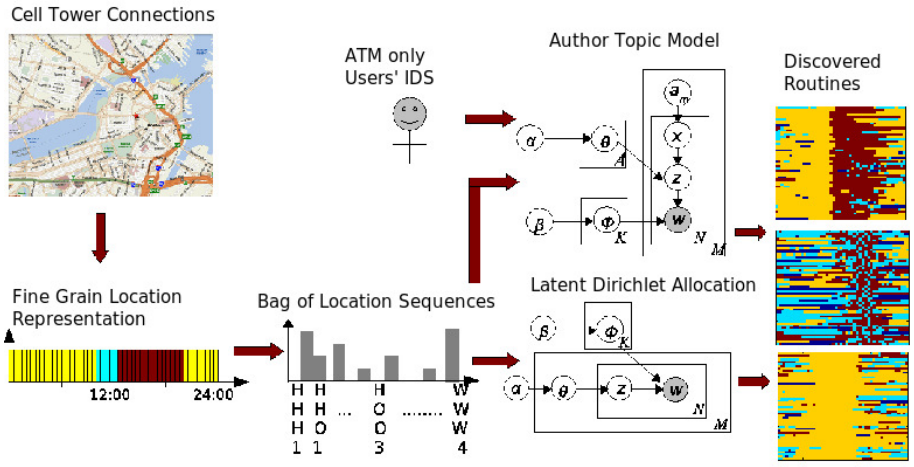


Figure 3.2. Block diagram of our methodology. User locations, given by cell tower connections, are transformed into bag of location sequences by first representing a user location as a fine grain location representation. This bag of location sequences is passed to Latent Dirichlet Allocation, in the first set of experiments, resulting in the discovery of routines. In the second set of experiments, the bag of location sequences and the user IDs are input to the Author-Topic Model, resulting in the discovery of routines and user patterns.

of an individual in terms of their locations obtained by cell tower connections and use this information to form a *bag of location sequences*. This bag representation was carefully designed to capture dynamics (i.e., location transitions) as well as both fine-grain (30 minute) and coarse-grain (several hour) time descriptions. Details of the method for bag construction are in Section 3.2. Overall, we make an analogy between the bag of location sequences for mobile data and a bag of words for text documents, where a location sequence is analogous to a text word, a day in the life of a person is analogous to a document, and a person is analogous to the author of a document. We use two models to discover routines. The first is Latent Dirichlet Allocation (LDA), illustrated in Figure 3.2, in which the input is the bag of location sequences. The second is based on the Author Topic Model (ATM), also visualized in Figure 3.2, in which user identity is input to the model in addition to the location sequences. The output of the models is a set of probability distributions over words and latent topics, capturing the dominating underlying routines in the dataset. We can then rank location sequences and days per topic, as well as users per topic in the case of ATM, and observe the routines discovered as topics.

Our proposed methodology based on a bag of location sequences structure is advantageous over [25] in that it contains both fine-grain and coarse-grain time considerations, which keeps into account transitions in location and is robust to variations in the data which may be due to noise or due to variations in the dataset, such as eating lunch at 11:30 am as opposed to 11:55 am. Further, due to our location sequence structure, we can discover routines characteristic of various intervals in the day. Our topic model methodology clearly defines a mechanism to rank users and days (with probabilities always greater than zero unlike Eagle's methodology), and with easily identifiable routines with semantic meanings which can be visualized comprehensibly over many of the discovered topics. Ranking allows us to see the raw data in a particular order (given by probabilities), giving structure to the data. This is true for both users and days and is useful for visualizing and structuring the data. We can perform several tasks with the discovered data, such as find users that go to work early, find groups of users that are at home during the day, or find users that turn their phones off in the morning. We also discover a varying sort of routines based on the ATM, a differing methodology which incorporates individual user identities into the model. With this, we can rank users and days, characterize users and day structures based on the number of topics composing most of the probability mass, and analyze an individual's daily life patterns over time. Finally, PCA is a traditional technique in pattern recognition [22] which cannot be easily extended for very large databases due to numerical linear algebra constraints (matrix inverse and determinant computation for large matrices becomes infeasible). The more modern techniques we are considering are state-of-the-art models and an active research domain in machine learning.

3.2 Bag Representations

We describe our bag representation design for location sequences. By constructing a bag representation to capture fine- and coarse-grain location, both can be encoded and can be viewed as analogous to words for text mining.

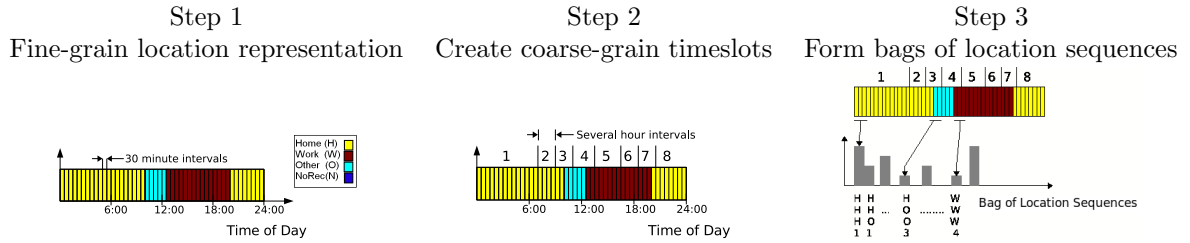


Figure 3.3. The construction of the bag of location sequences explained illustratively in a 3 step process. Step 1 is the division of a day into fine-grain 30-minute time intervals, which we call the fine-grain location representation. This first step includes associating a single location label of H, W, O, or N to each 30-minute time interval. Step 2 is division of a day into 8 unevenly distributed coarse-grain time intervals. Step 3 demonstrates the word construction. Three consecutive fine-grain location labels and the coarse-grain time interval are combined to form location sequences. The set of location sequences for a day forms the bag of location sequences.

3.2.1 Fine-Grain Location Representation

For a given individual in the dataset, there are entries for the set of cell towers users connect to, the start and end connection times. Over 32 000 towers are seen by all the users phones over the course of the year. Since we are interested to discover routines existing in the location data, we classify the cell towers into 3 semantic categories, removing the noise of the actual tower ID. As stated in Section 3.1, the categories were *home* (H), *work* (W), and *other* (O), representing towers which were the self-declared homes of users, work towers at MIT campus, and other towers, respectively. An initial set of labels was obtained from Nathan Eagle, one of the creators of Reality Mining. The list of W towers obtained from MIT was incomplete as several students never connected to any of those towers and thus were never considered to be at work. To resolve this issue, additional W labels were inferred from being in proximity to each person's computer; we did not consider being in proximity to one's laptop as being at work due to the mobile nature of the device. There was a fourth *no reception* (N) label, applied when there was no tower connection recorded for a user at a given time.

Following the labeling of cell towers into location categories, the day in the life of a user can then be expressed as a sequence of these location labels. The first step in forming our bag of location sequences is the construction of a fine-grain location representation, as illustrated in Figure 3.3, Step 1. We chose to a divide a day into fine-grain, 30-minute time intervals, resulting in 48 time blocks per day. We use 30-minute slots as many events of daily life are synchronized with half-hourly schedules and this does not result in vocabulary size explosion as discussed in the following section. For each block of time, we choose the location label which occurred for the longest duration, resulting in a single location label per timeslot. This is an important step as tower connections can be noisy and fluctuative. The result is a day of a user represented as a sequence of 48 location labels, visualized for the entire input dataset in Figure 3.1 (b).

3.2.2 Bag of Location Sequences

The bag of location sequences is built from the fine-grain location representation considering 8 coarse-grain timeslots in a day, as shown in Figure 3.3, Step 2. We divide a day as follows: 0-7 am (1), 7-9 am (2), 9-11 am (3), 11 am-2 pm (4), 2-5 pm (5), 5-7 pm (6), 7-9 pm (7), and 9-12 pm (8). The goal of these coarse-grain timeslots is to remove some of the potential noise due to minor time differences between daily routines. For example, if a user leaves the house at 7:30 am as opposed to 8 am, we want to capture the important feature of "leaving the house early in the morning" and not the minor time difference of this routine between days. The choice of the timeslots is also guided by common sense about daily activities (e.g., typical lunch times, working times, sleeping times).

The last step in building the bag of location sequences is the word construction, visualized in Figure 3.3, Step 3. A location sequence contains 3 consecutive location labels in the fine-grain representation, corresponding to 1.5 hour intervals, followed by one of the 8 timeslots in which it occurred. Thus a location sequence has 4 components, 3 location labels followed by a timeslot. We take overlapping 1.5 hour sets of labels to make a location sequence, so that if we had a pattern HHHOW in the interval 7 am-9:30 am, we would have for 7:30 am, 8 am, and 8:30 am, the following location sequences: HHH1, HHO1, and HOW1, where 1 indicates timeslot 1. Finally, the bag of location sequences is the histogram of the location sequences present in the day. In the analogy with text, a document is a day of a user and an author is a user.

3.3 Topic Models for Routine Discovery

As discussed in Chapter 2, topic models are powerful tools initially developed to characterize text documents, but can be extended to other collections of discrete data. They are probabilistic generative models that can be used to explain multinomial observations by unsupervised learning. Formally, the entity termed *word* is the basic unit of discrete data defined to be an item from a vocabulary of size V. A *document* is a sequence of N words. A *corpus* is a collection of M documents. There are K latent topics in the model, where K is a hyperparameter in the simpler topic models.

3.3.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) (Figure 3.4 (a)) is a generative model, introduced by [10], in which each document is modeled as a multinomial distribution of topics and each topic is modeled as a multinomial distribution of words. The generative process begins by choosing a distribution over topics $\mathbf{z} = (z_{1:K})$ for a given document. Given a distribution of topics for a document, words are generated by sampling topics from this distribution. The result is a vector of N words $\mathbf{w} = (w_{1:N})$ for a document.

LDA assumes a Dirichlet prior distribution on the topic mixture parameters θ and ϕ, to provide a complete generative model for documents. θ is an M x K matrix of document-specific mixture weights for the K topics, each drawn from a Dirichlet(α) prior, with hyperparameter α. ϕ is an V x K matrix of

(a) Latent Dirichlet Allocation (LDA) (b) The Author-Topic Model (ATM)

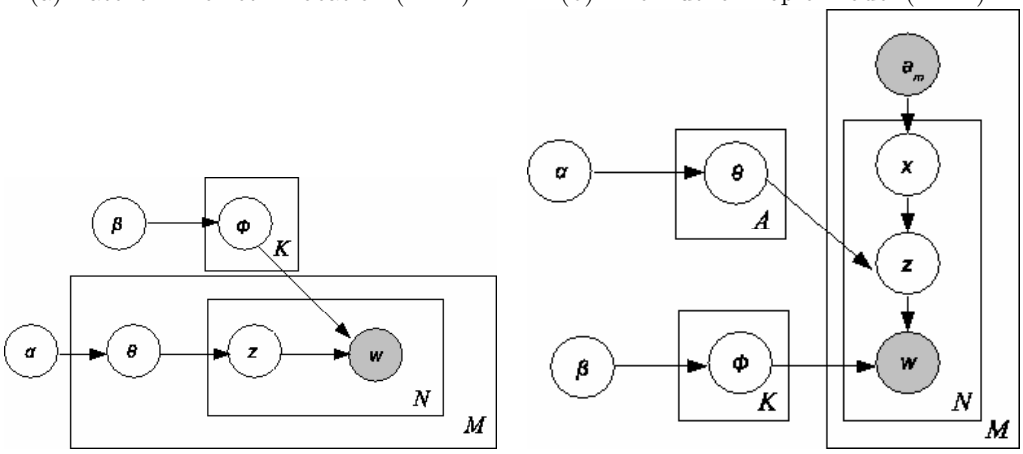


Figure 3.4. Graphical models of two probabilistic topic models (a) Latent Dirichlet Allocation (LDA) and (b) the Author Topic Model (ATM).

word-specific mixture weights over V vocabulary items for the K topics, each drawn from a Dirichlet(β) prior, with hyperparameter β.

The main objectives of LDA inference are to

1. find the probability of a word given each topic k, $p(w = t|z = k) \triangleq \phi_k^t$, and
2. find the probability of a topic given each document m, $p(z = k|d = m) \triangleq \theta_m^k$.

Several approximation techniques have been developed for inference and learning in the LDA model [10, 38]. In this work we adopt the Gibbs sampling approach [38].

For the LDA model visualized in Figure 3.4 (a), the following distributions hold:

$$p(\theta|\alpha) = p(\theta) \sim Dirichlet(\alpha) \tag{3.1}$$
$$p(\phi|\beta) = p(\phi) \sim Dirichlet(\beta) \tag{3.2}$$
$$p(z|\theta^{(d)}) \sim Multinomial(\theta^{(d)}) \tag{3.3}$$
$$p(w|z,\phi) \sim Multinomial(\phi^{(z)}) \tag{3.4}$$

where ϕ represents the word distribution for topic z, and $\theta^{(d)}$ represents the topic distribution for document d.

From the assumptions in equations 3.1-3.4, we obtain $p(w|z,\phi) = \prod_{k=1}^{K}\prod_{t=1}^{V}(\phi_k^t)^{n_k^t}$ where n_k^t is the number of times word t is assigned to topic k. n_k^t is also called the word-topic count and $n_k = \sum_{t=1}^{V} n_k^t$ is called the word-topic sum. We also obtain $p(z|\theta) = \prod_{m=1}^{M}\prod_{k=1}^{K}(\theta_m^k)^{n_m^k}$ where n_m^k is the number of times topic k occurs in document m. n_m^k is also called the topic-document count, and $n_m = \sum_{k=1}^{K} n_m^k$ is called the topic-document sum.

Further details of the Gibbs sampling for LDA model parameter estimation can be found in [38]. In practice, we can use the procedure summarized in Figure 3.5 to estimate the model parameters.

// GOAL: Given a training corpus, α, β, and K, estimate the parameters n_m^k and n_k^t from which we can determine the model parameters $\hat{\phi}_k^t$ and $\hat{\theta}_m^k$.

// Initialization
1) Initialize the count parameters, $n_m^k = 0$, $n_k^t = 0$.
2) **Iterate** over each word w in the corpus:
    3) Sample a topic k from $k \sim Mult(\frac{1}{K})$.
    4) Update the count parameters n_m^k, n_k^t as follows $n_m^k = n_m^k + 1$, $n_k^t = n_k^t + 1$.

// Run the chain
5) **Iterate** over a large number of iterations (e.g. 1000):
    6) **Iterate** over each word w in the corpus:
        7) Decrement the current word w and current word's topic assignment t counts as follows $n_m^k = n_m^k - 1$, $n_k^t = n_k^t - 1$.
        8) Sample a topic k from $p(z = k|\mathbf{z}_{\neg i}, \mathbf{w}) \propto \frac{n_k^t+\beta}{\sum_{t=1}^{V} n_k^t+\beta} \cdot \frac{n_m^k+\alpha}{\sum_{k=1}^{K} n_m^k+\alpha}$.
        9) Increment the new word/topic and topic/document counts as follows $n_m^k = n_m^k + 1$, $n_k^t = n_k^t + 1$.

// Compute model parameters
10) Estimate the unknown parameters as follows
$\hat{\phi}_k^t = \frac{n_k^t + \beta}{n_k + V\beta}$, and $\hat{\theta}_m^k = \frac{n_m^k + \alpha}{n_m + K\alpha}$, where $\hat{\phi}$ and $\hat{\theta}$ are the model parameter estimates,
$n_k = \sum_{t=1}^{V} n_k^t$, and $n_m = \sum_{k=1}^{K} n_m^k$.

Figure 3.5. Gibbs Sampling Algorithm for LDA.

3.3.2 Author-Topic Model

The Author-Topic model (ATM), introduced by [82] is also a generative model for documents that extends LDA to include authorship information. In ATM, each author is associated with a multinomial distribution over topics and each topic, like LDA, is associated with a multinomial distribution over words. By modeling the interests of authors, it becomes possible to establish what topics an author writes about, which authors are likely to have written documents similar to an observed document, and which authors produce similar work.

For ATM, each word in a document is associated with two latent variables, an author, x, and a topic, z. The graphical model in Figure 3.4 (b) illustrates the process. The set of authors of document m is defined as a_m, where $A = |a_m|$ is the number of authors who generated the documents in the corpus. Furthermore, x indicates the author responsible for a given word, chosen from a_m. In this model, ϕ denotes the V x K matrix of word-topic distributions, with a multinomial distribution over V vocabulary items for each of K topics drawn independently from a Dirichlet(β) prior. θ is the A x K matrix of author specific mixture weights for these K topics, each drawn from a Dirichlet(α) prior.

The main objectives of ATM inference are to

1. find the probability of generating word t from topic k, ϕ_k^t and
2. find the probability of assigning topic k to a word generated by author a, θ_a^k.

For the ATM model visualized in Figure 3.4 (b), the following distributions hold:

$$p(\theta|\alpha) = p(\theta) \sim Dirichlet(\alpha) \tag{3.5}$$
$$p(\phi|\beta) = p(\phi) \sim Dirichlet(\beta) \tag{3.6}$$
$$p(z|x, \theta^{(x)}) \sim Multinomial(\theta^{(x)}) \tag{3.7}$$
$$p(w|z, \phi^{(z)}) \sim Multinomial(\phi^{(z)}) \tag{3.8}$$
$$p(x|a_m) \sim Uniform(a_m) \tag{3.9}$$

where $\theta^{(x)}$ represents the topic distribution for authors x.

Parameter estimation in ATM with Gibbs sampling is based on the procedure in Figure 3.6. For Gibbs sampling, the joint conditional probability distribution defined in Step 9 of Figure 3.6 is used [82], where the word-topic count, n_k^t is the number of times word t is assigned to topic k and $n_k = \{n_k^t\}_{t=1}^V$ is the word-topic sum. The topic-author count, n_a^k, is the number of times author a is assigned to topic k, and $n_a = \{n_a^k\}_{k=1}^K$ is the topic-author sum.

3.3.3 Perplexity and Model Selection

Perplexity is a common measure in text modeling of the ability of a model to generalize to unseen data [42]. We use perplexity as a measure to determine how well the topic model generalizes to a set of unseen documents given topics trained on a training corpus. It is defined as the reciprocal geometric mean of the likelihood of a test corpus given a model,

$$Perplexity = exp[-\frac{\sum_{m=1}^{M} \log p(w_m|\mathcal{M})}{\sum_{m=1}^{M} N_m}], \tag{3.10}$$

where N_m is the length of document m, $\mathcal{M}$ is the model, w_m are the set of unseen words in document m, and M are the number of documents in the test set.

In order to compute perplexity from a set of previously unseen documents, we:

1. Divide the entire corpus into two groups, training and test sets. We randomly chose proportions of 90% training and 10% test documents.
2. Run the inference algorithm on the training corpus.
3. Run the inference algorithm on the test corpus, but "shift" the topic weights according to those obtained in Step 2 (training phase). More specifically, sample the topic/word and topic/document counts of the test corpus, but add the topic/word count of the training corpus to β before sampling. In step 8 of Figure 3.5, β can be replaced by $n_{k_{train}}^t + \beta$, where $n_{k_{train}}^t$ are the topic/word counts from the training phase.

// GOAL: Given a training corpus, α, β, and K, estimate the parameters n_a^k and n_k^t from which we can determine the model parameters $\hat{\phi}_k^t$ and $\hat{\theta}_a^k$.

// Initialization
1) Initialize the count parameters, $n_a^k = 0$, $n_k^t = 0$.
2) **Iterate** over each word w in the corpus:
  3) Sample a topic k from $k \sim Mult(\frac{1}{K})$.
  4) Sample an author a from $a \sim Mult(\frac{1}{A_m})$ where A_m is the list of authors of document m.
  5) Update the count parameters n_a^k, n_k^t as follows $n_a^k = n_a^k + 1$, $n_k^t = n_k^t + 1$.

// Run the chain
6) **Iterate** over a large number of iterations (e.g. 1000):
  7) **Iterate** over each word w in the corpus:
    8) Decrement the current word t's topic k and author a assignment counts as follows $n_a^k = n_a^k - 1$, $n_k^t = n_k^t - 1$.
    9) Sample a topic k and author a assignment for the word from $p(x = a, z = k|\mathbf{w}, \mathbf{z}_{-i}, \mathbf{x}_{-i}, A, \alpha, \beta) \propto \frac{n_k^t + \beta}{n_k + V\beta} \cdot \frac{n_a^k + \alpha}{n_a + K\alpha}$.
    10) Increment the new word/topic and topic/author counts as follows $n_a^k = n_a^k + 1$, $n_k^t = n_k^t + 1$.

// Compute model parameters
11) Estimate the model parameters as follows
$\hat{\phi}_k^t = \frac{n_k^t + \beta}{n_k + V\beta}$, and $\hat{\theta}_a^k = \frac{n_a^k + \alpha}{n_a + K\alpha}$, where $n_k = \sum_{t=1}^{V} n_k^t$, and $n_a = \sum_{k=1}^{K} n_a^k$.

Figure 3.6. Gibbs Sampling Algorithm for ATM.

3.3.4 Topic Models for Routine Modeling

As stated before, to model human activities, we make an analogy between text documents and human location patterns. We replace words with location sequences, documents with days, topics with routines, and authors with users. The LDA model produces ϕ_t^k and θ_k^m, which represent the probability of location sequence t for each topic k, and the probability of topics k for each day m, respectively. Given these probability distributions, we can rank location sequences and days for each topic discovered, and determine routines which are discovered as topics.

The ATM model extends this interpretation, this time with the emphasis of determining distributions of topics over authors, or routines followed by users. The ATM model produces ϕ_t^k and θ_k^a, which represent the probability of location sequence t for each topic k, and the probability of topics k for each user a, respectively. Given these probability distributions, we can again rank location sequences for each topic discovered. Furthermore, with this methodology we can also rank topics for users, resulting in the discovery of routines followed by users.

Based on our method, we set out to answer the following questions:

– What type of topics do LDA and ATM discover?

- Are there specific activity patterns occurring on weekends versus weekdays?
- How do the topics discovered characterize the set of days and users in the dataset?
- Does the entropy of a user's location-routines have a meaning?
- How does the proposed method compare to classic clustering methods?
- Can the topic model methodology find changes in a user's daily location routines or discover meaningful groups?

We provide answers to these questions in the following section.

3.4 Experiments and Results

In this section, we present our results to the questions mentioned above. First we present the data used and describe the experiments used for model selection. We present the results of human location-driven activities from LDA and ATM. We then investigate daily patterns, compare our method to clustering, investigate users in terms of their location-entropy, and use the topic model method to discover groups of users' routines.

3.4.1 Data

As summarized in previous sections, in the Reality Mining dataset [26], the activities of 97 subjects were recorded by mobile phones over 491 consecutive days of data recording (01.01.2004 to 05.05.2005). More precisely, 25 of the students in the dataset are labeled as Sloan business and the remaining 72 individuals are students and staff from the Media Lab. For the experiments, we removed days which were entirely N (no reception), since they contained no useful information. The resulting dataset is still massive, amounting to 10 118 days, and over 242 800 hours of data.

3.4.2 Model Selection for LDA

We use perplexity as a measure to determine the optimal number of latent topics, K. We computed perplexity for LDA using K values from 20 to 500 with increments of 20. For all values of K, initialization was followed by 1000 iterations of the Gibbs sampling algorithm. The perplexity is plot over the number of latent topics in Figure 3.7. A drop in perplexity occurs at approximately $K = 200$ topics, after which the perplexity stabilizes. We choose $K = 200$ as the number of latent topics for the remaining experiments.

3.4.3 Routines Discovered with LDA

The LDA model successfully finds latent topics over all users and days, and contains the dominating location routines. The unsupervised discovery of location-driven routines revealed different types of patterns, assigning intervals of days with different patterns to various topics with a probability measure. To illustrate the routines discovered, for each topic we rank the 4 most probable location sequences,

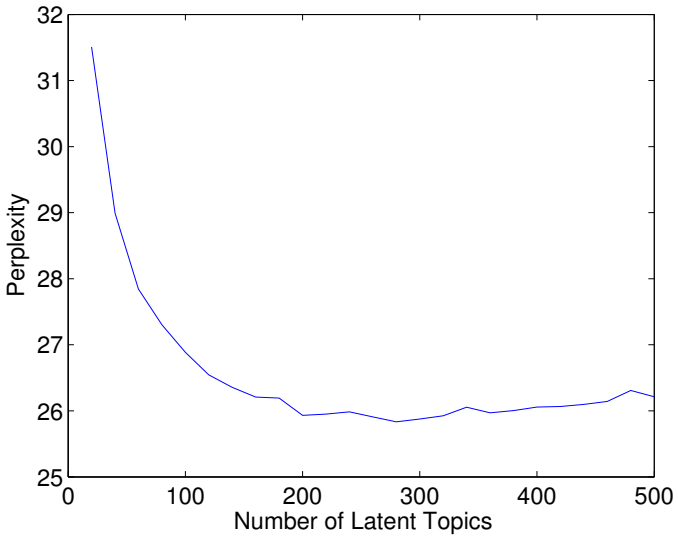


Figure 3.7. Perplexity plot as a function of the number of latent topics, K. At $K = 200$, the perplexity mostly stabilizes to a low value.

ranked by $p(w|z)$, and show them in tables. We also rank the 50 most probable days, ranked by $p(z|d)$, and visualize them in plots.

In Table 3.1, we illustrate various types of work routines exhibited by listing the top location sequences with the corresponding topics' visualization of top days, ranked by $p(z|d)$.

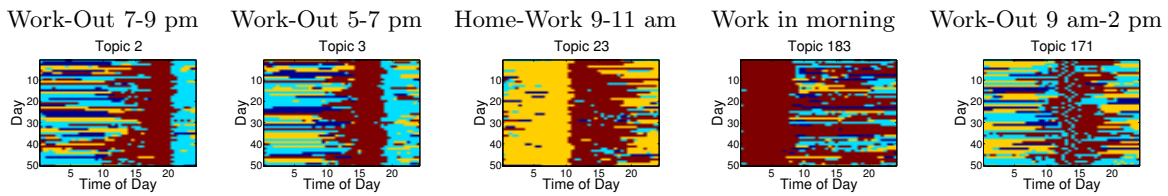
Some interesting results are the following:

- Topics 2 and 3 in Table 3.1 capture "going from work to out in the evening" routines, at different time intervals. The most probable words for topic 2 are WWW6, which is being at work in timeslot 6 (5-7 pm) followed by going from work to out in timeslot 7 (7-9 pm) WOO7, WWO7. Topic 3 contains very similar top words, but in one timeslot sooner: it is characterized by being at work in timeslot 5 (2-5 pm) followed by going from work to out in timeslot 6 (5-7 pm). Beneath the table, we visualize the top days for those topics, and can see that the days in topic 3 contain a work - out transition at an earlier interval than in topic 2.
- Topic 23 captures a "going from home to work" routine between 9-11 am. The most probable words are "at home before 9 am", followed by HWW3, HHW3, which represent "going from home to work" transitions in timeslot 3 (9-11 am).
- Topic 183 captured "at work early in the morning", with the most probable words being WWW1 and WWW2 followed by transitions around 7-9 am.
- Topic 171 illustrates a "work to out fluctuation in the early afternoon" with top words containing work to out fluctuations in timeslots 3 and 4 (9 am-2 pm).

Note that in all these topics, the top few words account for over 90% of the probability mass, which suggests that the topics are discriminant of very characteristic patterns despite the inherent noise present in most days' data. This is possible due to the relatively large number of topics we use.

Table 3.1. Work Routines: The table lists the 4 most probable location sequences ranked by $p(w|z)$ for topics 2, 3, 13, 23, 183, and 171. The visualizations beneath illustrate the corresponding 50 most probable days ranked by $p(z|d)$, entitled with the topic number and the semantic work routine (given by manual inspection).

Topic 2 - LDA		**Topic 3 - LDA**		**Topic 23 - LDA**		**Topic 183 - LDA**		**Topic 171 - LDA**	
Word	$p(w\|z)$	**Word**	$p(w\|z)$	**Word**	$p(w\|z)$	**Word**	$p(w\|z)$	**Word**	$p(w\|z)$
W W W 6	0.548	W W W 5	0.462	H H H 2	0.528	W W W 1	0.920	W W O 4	0.300
W O O 7	0.212	W W O 6	0.255	H W W 3	0.212	W W W 2	0.020	O W W 4	0.290
W W O 7	0.196	W O O 6	0.231	H H W 3	0.201	W W O 2	0.013	W O W 4	0.273
O O W 3	0.003	O W W 4	0.003	H H H 3	0.022	W O O 2	0.008	O O W 3	0.046



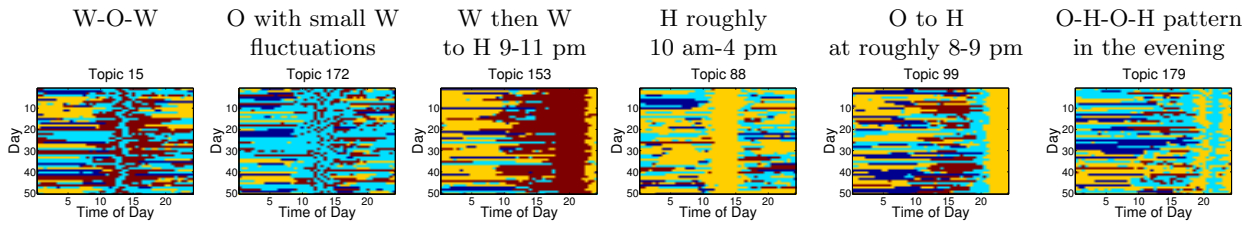
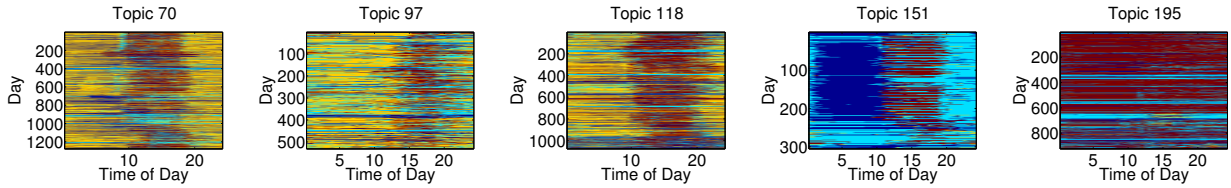


Figure 3.8. A small subset of the routines discovered visualized for the top 50 days for each topic. The corresponding routine name is displayed above the discovered topics.

Table 3.2. A selection of discovered ATM routines. Top location sequences are listed for selected topics, ranked by $p(w|z)$, as well as top users for these topics, ranked by $p(z|a)$. Beneath are plots for the top users' days for given topics illustrating the routines discovered.

Topic 70 - ATM		**Topic 97 - ATM**		**Topic 118 - ATM**		**Topic 151 - ATM**		**Topic 195 - ATM**	
Word	$p(w\|z)$	**Word**	$p(w\|z)$	**Word**	$p(w\|z)$	**Word**	$p(w\|z)$	**Word**	$p(w\|z)$
H H H 1	0.346	H H H 1	0.289	H H H 1	0.221	N N N 1	0.193	W W W 1	0.332
H H H 8	0.151	H H H 4	0.159	W W W 5	0.16	W W W 5	0.175	W W W 5	0.121
H H H 7	0.128	H H H 3	0.14	W W W 6	0.14	W W W 4	0.125	W W W 4	0.12
H H H 6	0.064	H H H 2	0.113	W W W 4	0.113	N N N 2	0.116	W W W 2	0.099
User	$p(z\|a)$	**User**	$p(z\|a)$	**User**	$p(z\|a)$	**User**	$p(z\|a)$	**User**	$p(z\|a)$
95	0.286	62	0.348	54	0.234	14	0.320	26	0.639
11	0.226	57	0.209	29	0.219	43	0.100	27	0.618
15	0.213	63	0.163	10	0.160	78	0.089	58	0.578
39	0.205	75	0.156	85	0.153	8	0.081	24	0.526



Other routines discovered are visualized in Figure 3.8 with their corresponding manually assigned labels as the title. Note that these selected routines are just a few of the many meaningful topics discovered.

- Topics 15 captures a work - out - work routine which could correspond to a lunch break.
- Topic 172 captures being out most of the day with very short work fluctuations occurring between 10 am-3 pm.
- Topic 153 captures working non-stop for at least 4 hours, then going home at approximately 10 pm.
- Topic 88 captures home roughly 10 am-3 pm.
- Topic 99 captures out for a few hours at roughly 8 pm, then arriving home at around 9 pm and staying home for the entire evening.
- Topic 179 captures an out-home-out-home routine, with each location occurring for a few hours in the evening.

3.4.4 Routines Discovered with ATM

For the ATM we use the same model parameters as those for LDA. Specifically, $K = 200$, $\beta = 0.1$, $\alpha = 50/K$, and we run 1000 iterations of the Gibbs sampling algorithm. The results obtained by the ATM differ from those obtained by LDA. With the ATM, the routines capture users' routines, simultaneously taking into account users' identities and daily location routines. In contrast, the LDA model captures routines from the days in the dataset, disregarding users' identities.

In Table 3.2, selected ATM results are listed. We include the top 4 location sequences for selected topics, ranked by the probability of a word given the topic, $p(w|z)$. We also include the top 4 users for the selected topics, ranked by the probability of the topic given a user, $p(z|a)$. Beneath the table, the plots entitled "Topic x", display all the days of users for which $p(z|a) > T_a$, where $T_a = 0.1$, ranked by users. We pick a selection of 5 out of 200 topics to demonstrate the routines obtained. Note that each user has a different number of recorded days in the dataset, and each topic has differing number of users with $p(z|a) > T_a$, which explains the varying number of days plot for each topic.

- In Topic 70, the top words are "being at home in the early morning (HHH1) and evening from 5 pm onwards" (HHH6, HHH7, HHH8). Users whose daily lives most often evolve around this routine are users 95, 11, 15 and 39 who characterize this topic with similar probabilities.
- Topic 97 discovered a "being at home early in the day" (HHH in timeslots 1-4). Users 62, 57, 63 and 75 display this routine most frequently, though not everyday, as seen by the lower $p(z|a)$ for users 63 and 75. In the corresponding plot, we can see a general "being at home in the mornings and afternoons" routine, though not everyday.
- Topic 118 found a "being at home" in the morning routine (HHH1) co-occurring with "being at work 11 am-7 pm" (WWW4, WWW5, WWW6). Users 54, 29, 10, and 85 most frequently follow this daily life pattern.
- Topic 151 captures "no reception in the morning" (NNN1 and NNN2) co-occurring with "being at work in the afternoon" (WWW4, WWW5). Users 14 and 43 most strongly follow this routine.
- Topic 195 discovered a "being at work throughout the day" routine (WWW in timeslots 1,2,4,5), which is very frequently followed by users 26, 27, 58, and 24, seen by their high $p(z|a)$ and their daily lives visualized in the figure. This could potentially be the discovery of "users who live on campus".

Comparing the routines obtained with LDA and ATM, we note that the ATM produces topics with top words that do not account for the probability mass as strongly as they do in LDA. Furthermore, note that none of the top users shown for the topics are the same. This suggests that the ATM is preferring certain users versus others for these topics. Also, note that for some topics, users are very characteristic (high $p(z|a)$), while for other topics this is not the case. Overall, the ATM learned topics are more "general" than the ones with LDA, with the advantage of learning the user-topic distributions. We lose discrimination in the topics (routines discovered) with ATM but this is traded off for learning users' distributions.

3.4.5 Daily Patterns

Our methods allow to extract daily patterns that are meaningful according to the day type, and also seen as a mixture. We now discuss these two aspects.

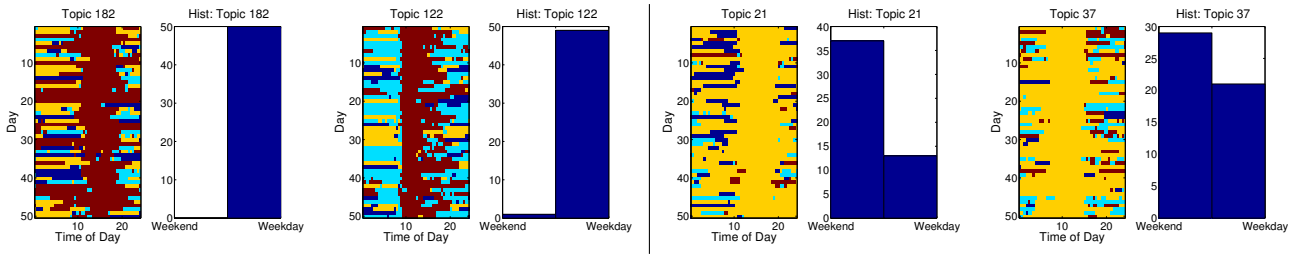


Figure 3.9. Weekend-dominant versus weekday-dominant routines discovered by LDA. The visualizations entitled "Topic x" show the top 50 days, ranked by $p(z|d)$, for topic x. The plots entitled "Hist:Topic x" are counts of whether the most probable days in topic x correspond to Weekends or Weekdays. It can be seen that the top 50 days for topics 182 and 122 almost entirely correspond to weekdays. The majority of the most probable days of topics 21 and 37 correspond to weekends.

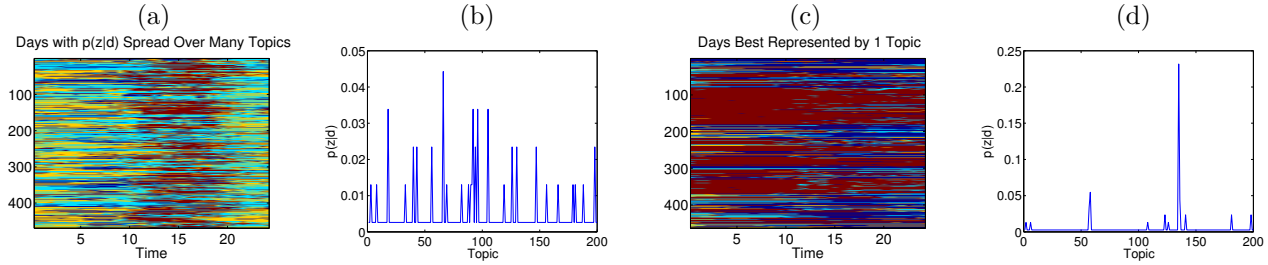


Figure 3.10. (a) Days which are described by many topics in LDA. (b) $p(z|d)$ plot for a given day which is described with low probability by many topics. (c) Days which are well described by a single topic. (d) $p(z|d)$ plot for a given day which is well described by a single topic.

Weekend and Weekday-Like Routines

On a weekly level, some trends characteristic of weekends versus weekdays appeared with the routines discovered by LDA. For example, topics 182 and 122, plot in Figure 3.9, demonstrate routines which dominated on weekdays and topics 21 and 37 demonstrate routines which tend to dominate on weekends. Each visualization of the most probable days per topic, entitled "Topic x" is followed by a histogram, entitled "Hist: Topic x", which counts whether the topic's 50 top days correspond to weekends or weekdays. We can see a "being at work" routine in topics 182 and 122 corresponds to weekday trends, and "being at home" during the day corresponds mostly to weekend behavior, though some weekdays also demonstrate this routine, perhaps corresponding to holidays or days off.

Days as Mixture of Topics

One fundamental question that arises is: how evident is the "mixture of topic" assumption in our data? Are days about one topic or more? Our LDA methodology allows us to find days which vary over many topics, and days which are best represented by a few topics. On one hand, by looking at days for which $p(z_i|d) <= T_L, \forall i = 1 : K$ where T_L is a small value (set here to 0.05 based on heuristic experiments), we can find days which are not highly probable for any topic and thus are distributed with low probabilities over many topics. These days are visualized in Figure 3.10 (a), and the topic distribution

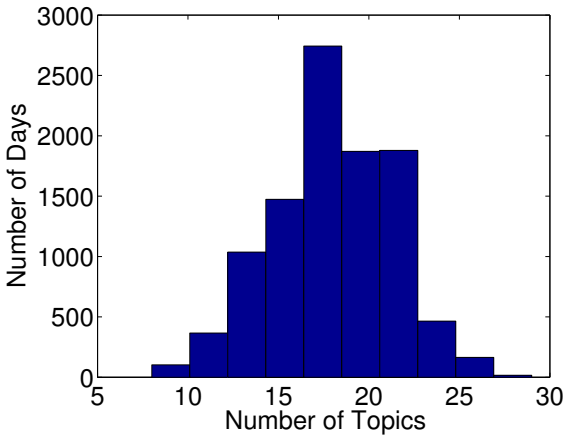


Figure 3.11. Histogram of number of 'dominating' topics per day for the LDA model.

given a particular day which is not highly probable for any topic is shown in Figure 3.10 (b). On the other hand, we also find days which are best represented by a few topics by collecting days for which $p(z_i|d) > T_H$, for a given i, where $T_H = 0.15$. These days are visualized in Figure 3.10 (c). In Figure 3.10 (d) we plot the topic distribution given a day which is well represented by a single topic. The thresholds T_L and T_H were picked in order to depict data on the order of 500 days. Comparing Figures 3.10 (a) and (c) we can differentiate between days following a rich set of routines and days lacking in variety in terms of location patterns. Those with highly varying routines, generally require more topics to capture their structure. The results here are highly dependent on the chosen value for K, the number of topics in LDA and ATM.

In Figure 3.11, we show a histogram of the number of "dominating" topics per day. We compute the number of topics composing at least 50% of the probability mass of each day in the study, and plot a histogram of the results. In general, all days are well described by fewer than 30 topics. Thus, at most 15% (30/200) of the topics can describe the probability mass of any day in the dataset. On the lower end of the histogram, very few days are described by less than 10 topics (35 days, or 0.3% of the days in the dataset). The same can be observed for high number of topics, very few days require 25 or more topics to be well defined (180 days, or 1.8% of the days in the dataset). The average number of topics in the study is 18 topics. Therefore, even though people typically follow very routine daily lifestyles, as found in [37], their daily location routines are true mixtures, involving a mixture of around 20 topics on average to define over 50% of the probability mass of the day.

3.4.6 Topics versus Clusters

One basic question is whether the topic model discovers groups of days different than classical clustering algorithms would. To investigate this, we compare the results of routine discovery from the k-means clustering algorithm [22] to those obtained by LDA. For this task we run k-means with 50 clusters and

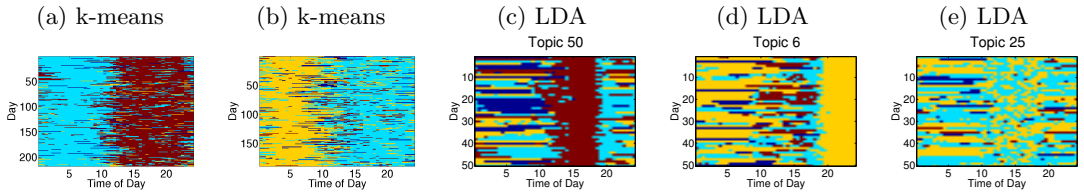


Figure 3.12. K-means clustering versus LDA topic discovery. (a) and (b) illustrate two typical clusters obtained by k-means. (c)-(e) illustrate three topics obtained by LDA. K-means clustering discovers very general routine discovery, occurring over the entire day. Topic discovery results in the probabilistic ranking of discriminative words. These discriminative words result in the determination of routines occurring dominantly over parts of the day. Further, transitional patterns, such as the home-out fluctuations in (e), can be found with LDA, but not with k-means.

compare the results to LDA with $K = 50$ topics. The input to k-means is the fine-grain location representation (in binary). Both algorithms are initialized randomly. The results are illustrated in Figure 3.12. Results are presented for a small number of topics for simplicity of visualization and analysis. We observe that k-means finds very general routines occurring 'broadly' over the entire day. In contrast, LDA finds topics with patterns occurring over parts of the day, as well as days with specific transition patterns occurring at a given time, such as those shown in Figure 3.12 (c) and (d). Furthermore, LDA discovers several routines such as the one visualized in Figure 3.12 (e), where alternating locations occur for varying time durations, which are not found by k-means. They are discovered with LDA due to the exchangeability of words assumption [10], which cannot likely be found using basic clustering techniques which take the exact occurrence of labels (here location) into account for comparison between data vectors (here days). More generally, two advantages of topic models over traditional clustering methods are: soft clustering of days, and meaningful word distributions as the representation of topics. Concretely in our work, the LDA model results in probabilistic distributions of days given all topics whereas k-means assigns only one cluster per day, and discriminative location sequences per topic characterizing human routines. This information is very useful as we know the precise location transitions which characterize the human routine as well as the timestamp, giving the routine's time details.

3.4.7 Modeling Individual Users with Topics

LDA-Based User Analysis

We can also examine the topic distributions over users with LDA. For each user x's day d_x, we count the topics for which the ranked probability of the topics given the day, $p(z|d_x)$ is greater than T_L, aggregate them for all the user's days, and illustrate them in the histogram entitled "User x Dominant Topics" in Figure 3.13. Some users' days are expressed well by a few topics, other users have a rich set of varying routines which are expressed as a mixture of many topics. For example, noting the varying y-axis scales, user 14 has a very high probability of a few topics for most days, whereas users 29, 95, and 97's days are expressed as a mixture over many topics. We plot the users' location data in the plots entitled "User x Data". Each user has a different number of days ($y-$axis), since they have varying number of

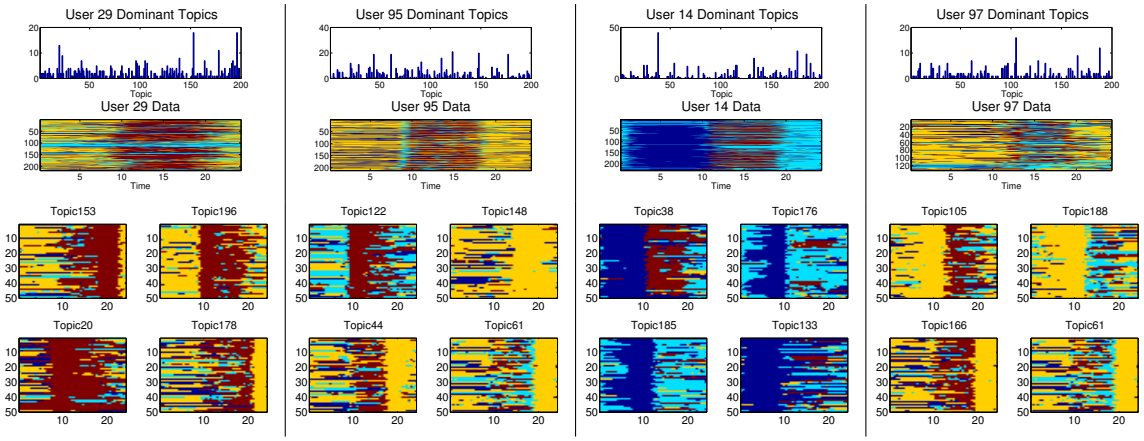


Figure 3.13. Individual user analysis. The histograms "User x Dominant Topics" demonstrate dominant topics for users x. Plots "User x Data" corresponds to the raw input location data of user x. The four topics below are the four dominating routines for user x.

days after removing fully *no reception* days. Beneath the users' days are the four topics which dominate the given users' daily activities. For instance, the four topics dominating user 29's daily routines are topics 153, 196, 20, and 178. User 29's dominating routines are "being at work" routines, as well as "being at work late in the evening". Looking at 'User 29 Data', we can confirm that user 29 does work a lot, especially late in the evening. User 29's daily activities are thus a mixture over several topics, as can be seen by the histogram "User 29 Dominant Topics". User 95's most common routines are "arriving to work before 10 am from an O location", and "being at home in the afternoons and evenings". Looking at user 95's location data, we can see this user is at home in the morning then goes to an O location, perhaps for breakfast or the gym, then goes to work and home in the evenings. User 14 mostly has no reception in the mornings, followed by being at work during the day, as seen by the dominant topic 38 dominating most of his/her daily activities. The user is often out in the evenings. It appears that this user's home label is missing, and she/he either turns the phone off while sleeping, or loses reception early in the morning. Overall, user 97's routines are predominantly "at home in the morning and evenings". User 14's dominant topics are less of a mixture over several topics than users 29, 97 and 95.

ATM-Based Analysis

We can also analyze individual user's daily structure with the topics discovered by ATM. The plots in Figure 3.14 illustrate how well users' daily routines are described by the topics discovered by the ATM. In Figure 3.14 (a) we plot the number of dominating topics composing 70% of the probability mass for each user. Most users' daily routines are described well by less than 17 topics. Users with no data are not considered. Like with LDA, some users are well characterized by a few topics, others require more. In Figure 3.14 (b) we plot 2 individual users that vary in their daily routine distributions over topics. User 14 (also discussed for the LDA case) is well characterized by 2 topics, whereas user 65 is characterized by a mixture of many topics. In Figure 3.14 (c) and (d) we plot the days of 4 users whose daily life

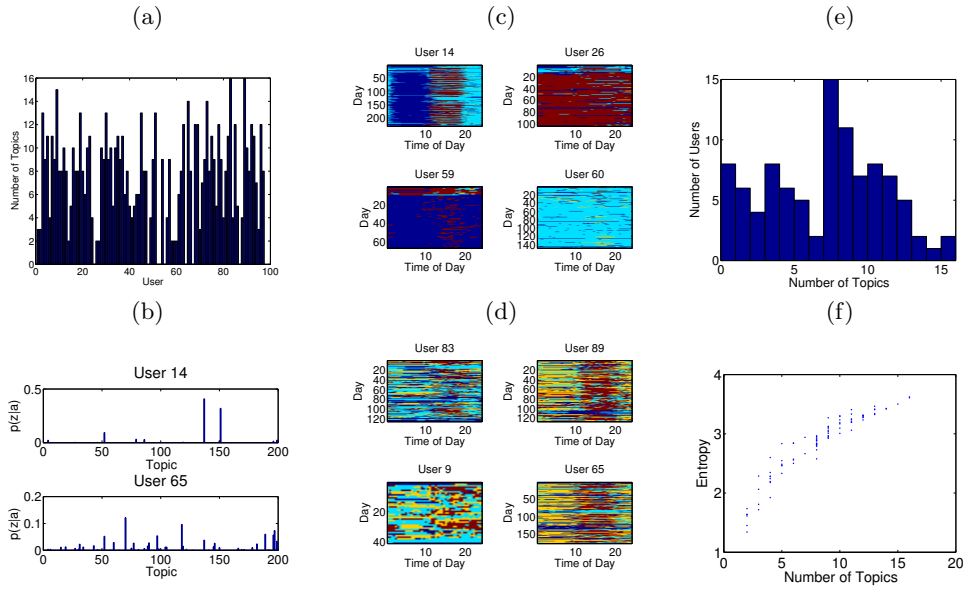


Figure 3.14. ATM Results. (a) Plot of the number of dominating topics for each user. Most users' daily routines are described well by less than 20 topics. (b) Topic distribution for users 14 and 65. User 14 is well characterized by 2 topics, whereas user 65 is characterized by a mixture of topics. (c) The days of 4 users whose daily life patterns are described well by a mixture of 2 topics. (d) The days of 4 users who are characterized by a mixture of many topics. (e) A histogram of the number of users as a function of the number of topics. (f) Number of topics plot as a function of entropy for each user.

patterns are described well by a mixture of 2 topics, and 4 users whose daily routines are described by a mixture over many topics. Visible differences between these users' lifestyles are that users 14, 26, 59, and 60 follow very regular, non-varying lifestyles, which are captured well by a few topics, whereas users 83, 89, 9, and 65 have varying daily routines. We also plot a histogram of the number of users whose lives are characterized by various number of topics in Figure 3.14 (e). According to the ATM analysis, many users can actually be well characterized by few topics. 10 users can be well characterized by fewer than 4 topics, and 18 users by fewer than 5 topics, demonstrating that a significant portion of users have very repetitive non-varying lives. Fewer users have more highly varying lifestyles, as seen by the higher end of the histogram. 11 users can be well described by more than 12 topics. In Figure 3.14 (f) we plot the entropy for each user, computed on the topic distribution, as a function of the number of dominating topics. Each data point represents an individual user. We can see that the number of topics as a function of entropy is about linear, suggesting that the number of dominating topics is indeed a good measure of user variation in daily activities.

3.4.8 Finding Variations in Individual Lives Over Time

In order to find variations in a user's daily routines over time, we compute the Bhattacharya distance between consecutive days of a user, $BD = \sqrt{1 - \sum_{z \in K} \sqrt{p(z)q(z)}}$, where $p(z) = p(z_k|d_i)$ and $q(z) = p(z_k|d_{i+1})$, where d_i and d_{i+1} indicate consecutive days. The Bhattacharya distance measures the

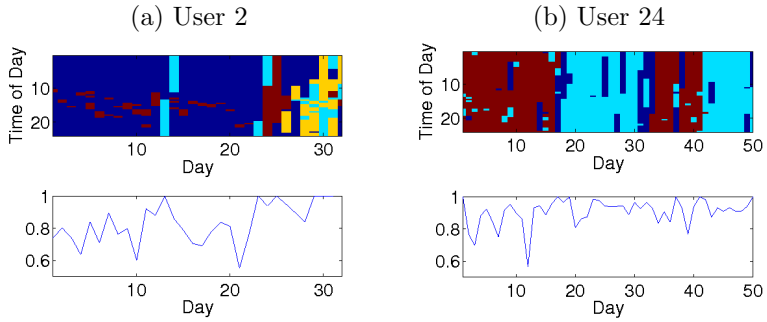


Figure 3.15. The difference between routines over consecutive days plot for (a) User 2 and (b) User 24. The top plots visualize several days with the time of day in hours ($y-$axis) vs. consecutive days ($x-$axis). The bottom plots correspond to Bhattacharya distance between pairs of consecutive days.

similarity of two discrete probability distributions. The more similar two probability distributions are, the closer the sum of products term in BD will be to 1. The smaller the overall BD term, the more similar two probability distributions, with a minimum value of 0. If two probability distributions differ significantly, the sum of product terms will be smaller, and the resulting BD expression will approach its maximum value of 1.

This measure proves to be useful in finding changes in a user's routines over time, as illustrated for users 2 and 24 in Figure 3.15. The bottom figures are the Bhattacharya distance plot as a function of day. The top figures are the set of days for both users, corresponding to the days in the Bhattacharya distance plots above. Note that here days are on the $x-$axis and time of day on the $y-$axis. Figure 3.15 (a) shows that at days 13-14, there is a change in the user's activities where they go from "N in the morning and evening" to "being out". This change is captured in the Bhattacharya distance, where the first peak to 1 occurs. The second peak occurs at day 23, where there is again a large variation in the user's routines, this time consisting of 'W' routines for a few days followed by 'H' routines for some days. User 24's changes in daily routines are also captured by the Bhattacharya distance, where the measure peaks to 1 at day 17, 19, 37, and 42.

After closer inspection, we noted that this measure is sometimes sensitive to days for which large variations do not occur. This is due to several topics accounting for similar routines, more likely to occur given the relatively large number of topics. For example, there are more than one topic displaying "being at work in the morning" or "being out during the day". Given the stochastic nature of Gibbs' sampling, two very similar days could have differing topic distributions. Thus, there are certain days which are very similar in structure but have larger than expected Bhattacharya distance measures. This could be a result of the potential inadequacy of the perplexity measure in model selection, which is discussed further in Section 3.5.1.

Table 3.3. Users that follow the specific routines of "going to work early", "working late then going home", and "turning off their phones (or no reception) in the evening. We list users that follow these routines more than usual.

Topic 196 - LDA W early from H			**Topic 122 - LDA W early from O**			**Topic 153 - LDA Work late then H**			**Topic 104 - LDA Phone off or N in evening**		
User	**Word**	$p(w\|z)$	**User**	**Word**	$p(w\|z)$	**User**	**Word**	$p(w\|z)$	**User**	**Word**	$p(w\|z)$
29	WWW3	0.61	95	WWW3	0.59	29	WWW7	0.44	90	NNN8	0.65
4	HHW2	0.17	41	WWW4	0.15	10	WWW6	0.32	94	NNN7	0.09
69	HWW2	0.12	13	OWW2	0.11	78	WHH8	0.10	42	ONN7	0.05
10	HWW3	0.04		OOW2	0.09		WWH8	0.10	37	OON7	0.04
	WWW2	0.01		HOW2	0.02		WHO8	0.01		HNN7	0.04

3.4.9 Extracting Group-Level Routines

LDA-Based Analysis

Using the LDA methodology, we can determine users that exhibit certain routines such as 'working early'. We use LDA to rank top days for these routines, and then count the number of times these days are highly ranked for each user. The individuals displaying these routines the most are listed as Users in Table 3.3.

As three examples, we can first find users that "go to work early". There are two topics that display this routine, one of them displaying "going to work from home" (Topic 196) and the other one "going to work from other location" (Topic 122). In both cases the user is at work by timeslot 3 (9 am-11 am). Users 29, 4, 69, and 10 arrive to work early from home, as seen in Topic 196 in Table 3.3. Users 95, 41, and 13 arrive to work early from O. We can also find users that work very late before going home. Topic 153 dominates for users 29, 10, and 78 who often stay at work late (5-9 pm) and arrive home after 9 pm. Finally, we discover users that possibly turn off their phones (have no reception) in the evenings. Users 90, 94, 42, and 37 often have no reception after 7 pm.

ATM-Based Analysis

While groups of users that share a routine can be discovered by LDA, ATM is in general better suited for this task. With the ATM methodology, certain topics exhibit routines that dominate for mostly business students. In Table 3.4, we display 4 of the routines that were discovered for business students. We list the top users along with their student types. We also display the top location sequences for these topics.

- Topic 104 dominates for users 38, 50, 96, and 82, three of which are Sloan students. These students tend to go out from home late in the evenings in timeslot 8 (9-12 pm).
- Topic 139 is dominant for users 12, 82, 66, and 69, three of which are Sloan students. These users go from home to work in timeslot 7 (7-9 pm) or from work to home in timeslot 6 (5-7 pm).
- Topic 145 occurs mostly for users 1, 84, 50, and 42, three of which are Sloan students. These users are often at O locations throughout the day.

Table 3.4. Routines discovered by ATM showing business student activities. Displayed topics are discovered to be dominating for business (Sloan) students.

Topic 104 - ATM		**Topic 139 - ATM**		**Topic 145 - ATM**		**Topic 146 - ATM**	
User & Type	$P(z\|a)$	**User & Type**	$P(z\|a)$	**User & Type**	$p(z\|a)$	**User & Type**	$p(z\|a)$
38 Sloan	0.040	12 Sloan	0.022	1 No label	0.187	4 Sloan	0.130
50 Sloan	0.036	82 Sloan	0.011	84 Sloan	0.125	73 Sloan	0.100
96 student	0.033	66 Sloan	0.011	50 Sloan	0.111	36 Sloan	0.091
82 Sloan	0.027	69 newgrad	0.009	42 Sloan	0.088	79 Sloan	0.085
Word	$p(w\|z)$	**Word**	$p(w\|z)$	**Word**	$p(w\|z)$	**Word**	$p(w\|z)$
H H H 8	0.115	H H W 7	0.164	O O O 5	0.258	H H H 5	0.158
H H O 8	0.099	H W W 7	0.135	O O O 4	0.241	H H H 1	0.130
H H H 7	0.097	W H H 6	0.119	O O O 3	0.149	H H H 4	0.127
H O O 8	0.079	O W W 3	0.052	O O O 6	0.084	N N N 4	0.100

– Topic 146 occurs dominantly for four business students. They are often at home in timeslots 4 and 5 (11 am-5 pm).

3.5 Conclusion

We have presented a novel framework for location-driven routine discovery using probabilistic topic models. Using a dataset collected by 97 users' mobile phones over a period of 16 months from the Reality Mining project, we successfully discover routines characteristic of days and individuals in the study in an unsupervised manner. We have proposed a method to represent location sequences, and incorporated this into the LDA and ATM topic models. The resulting distributions of words for latent topics, as well as topics given days, and topics given users, reveal hidden structure of routines which we use to perform varying tasks, including finding users or groups of users that display given routines, and determining times when certain events or changes in events occur. In the next chapter, we extend this work by investigating the proximity dataset obtained by Bluetooth in addition to location data and use both single and multi-modal cues to predict missing data and to classify day types and user types.

3.5.1 Limitations

While we have shown that many insights about routines can be obtained with our approaches, our work has some limitations. The first one is the scope of users for which data was collected. The users considered are MIT students and staff, and their daily routines are clearly not representative of the whole society. However, their daily routines are expected to be similar, for the most part, to other university students and staff as well as working professionals. Another limitation is the way we select the number of topics. For LDA, perplexity is used as a measure to determine performance on unseen data. However, perplexity is not a "perfect" measure for model selection, since similar resulting topics is not considered in the perplexity computation. Overall, perplexity is not perfect for model selection, though other ways of choosing model parameters are not much better, and the issue of model selection for topic models is

an active problem [11]. Finally, there is no objective evaluation of the task conducted here, but this can be addressed in the next chapter.

Chapter 4

Multimodal Framework for Routine Discovery and Prediction

There is relatively little work on the investigation of large-scale human data in terms of multimodality for human activity discovery. In this chapter we suggest that human proximity, obtained by mobile phone Bluetooth sensor data, can be integrated with human location data, obtained by mobile cell tower connections, to mine meaningful details about human activities. We propose an extension to the method of the previous chapter which integrates location with interaction obtained from proximity data. Some of the human activities discovered with our multimodal data representation include "going out from 7 pm-midnight alone" and "working from 11 am-5 pm with 3-5 other people", further finding that this activity dominantly occurs on specific days of the week. We also find dominant work patterns occurring on other days of the week. We further demonstrate the feasibility of the topic modeling framework to discover human routines to predict missing multimodal phone data on specific times of the day.

Two fundamental problems in this domain are how to *discover* recurrent patterns in a person's life from multimodal data like proximity, location, and motion, and how to *predict*, based on the knowledge of a person's routines, her most likely routines at any given time. On the one hand, pattern discovery via unsupervised learning is often a necessity, given the potentially large number of relevant routine patterns of an entire population and the huge amount of unlabeled data that can be recorded with a phone over time [25, 31]. On the other hand, predictions from aggregated user observations are, arguably, some of the most useful outcomes of routine modeling, by inferring both where and with whom a user would most likely be in the future (for anticipation) or would most likely have been in the past (for cases of missing data).

While recent works have started to analyze both problems from location or proximity data - discovery and prediction in [25], discovery in [31] - one aspect that has not been investigated in depth is the role of multimodal integration in large-scale routine analysis. More specifically, how does the joint use of multiple modalities enhance the understanding of a person's routines, and how can this be efficiently

represented and automatically inferred? Proximity to known people (as a coarse approximation of face-to-face interaction) adds a rich element of social context that is very useful to complement or disambiguate many situations in daily life. For instance, being at home alone or with a large group having a party represent entirely different social situations, that would be nevertheless identical from the sole perspective of location. Such finer descriptions of routines based on multiple cues are clearly important to characterize users and their habits.

This chapter is organized as follows. First we present the overall framework and the joint location-proximity representation. We then describe the modeling details including the proposed method for missing data prediction. The experiments and results are the presented in Section 4.3. In Section 4.4 we present a supervised approach for day type and student type classification.

Various parts of the work presented here was originally published in modified form in [29, 30, 32, 34]. Note the supervised task is presented in [29] and the prediction framework in [32], with an extended journal version in [34].

4.1 Multimodal Framework

We use the same data described in Chapter 3.1. Given a day in the life of a person in terms of where they go and the number of people within the group they are in proximity with, our goal is to discover the most commonly occurring routines. Further, we use the combined location and proximity routines discovered to predict missing location and proximity data. An overview of the method is visualized in Figure 4.1. We represent a day in the life of a user in terms of where they are over a 90-minute time interval as well as the number of people they are with during this time interval within the Reality Mining population, forming a joint location-proximity data representation, described next. This joint data representation is input to the Latent Dirichlet Allocation (LDA) model, from which human routines are discovered, representing common forms of social interactions which occur at varying locations.

4.1.1 Joint Location-Proximity Representation

The joint location-proximity data representation is based on the concatenation of data corresponding to users' location, proximity, and a timeslot indicating a coarse-grain measure of the time of day for which this data is measured. The details follow.

Location Representation: The location representation is the same as that described in Chapter 3.2. The 1.5-hour interval location sequences are overlapping, resulting in 48 x 1.5-hour 3-label location sequences in a day.

Proximity Representation: For proximity data, we use the Bluetooth readings to consider proximity with people in the Reality Mining group. Bluetooth can detect other similar devices located within a 10-meter radius, and so Bluetooth is a reasonable (although clearly imperfect) proxy for social interactions, though there are various sources of noise making it challenging to work with. On the one hand, we could expect that people actually interacting will often be sensed by Bluetooth but many cases of

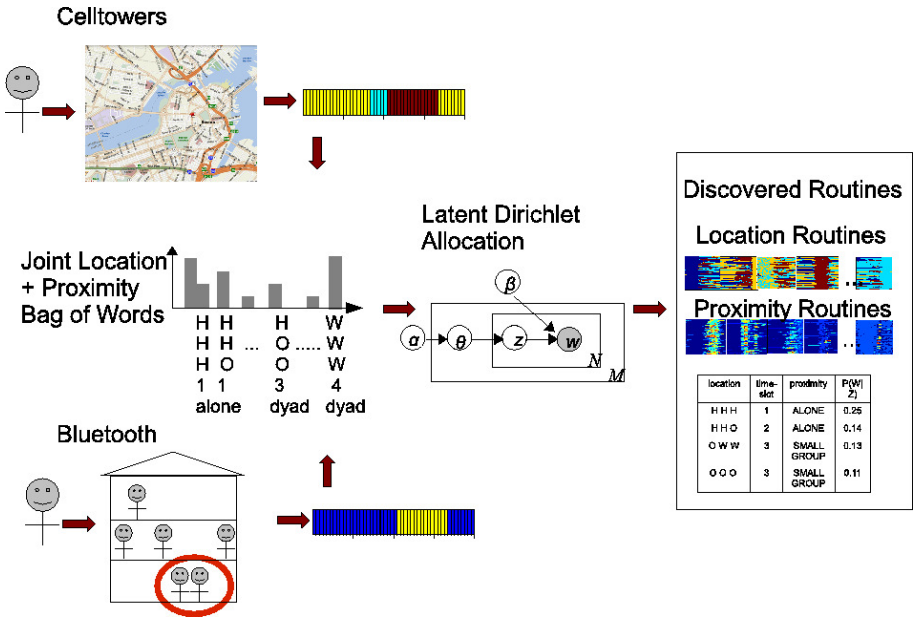


Figure 4.1. Overview diagram. The data captured by mobile phones (where a user is as well as with whom) is combined to form a joint location-proximity representation. After the multimodal data representation is transformed to a bag of words, Latent Dirichlet Allocation inference is applied to reveal latent topics (or discovered routines), corresponding to common user places and interactions. Each routine is characterized by its top multimodal words ranked by their probability.

nearby people who do not interact will be detected too. This is a limitation of the Bluetooth modality. Proximity in general could be considered as proximity to laptops, computers, and other people, but it is difficult to distinguish them from mobile phones. We quantize the number of proximate people into 4 prototypical groups: user is alone, dyad (1 person in proximity), small group (2-4 people in proximity), large group (5 or more people in proximity). The group sizes are motivated by research in social science that has traditionally analyzed dyads, small groups, and large groups as separate categories, as they present distinct dynamics [36].

Timeslot Division: As done in Chapter 3, each day is divided into 8 coarse-grain timeslots as follows: 0-7 am (1), 7-9 am (2), 9-11 am (3), 11 am-2 pm (4), 2-5 pm (5), 5-7 pm (6), 7-9 pm (7), 9-12 pm (8). These timeslots were chosen to capture common events in daily life, such as lunch time, dinner time, or morning and afternoon work times. Other time intervals could equally be used to capture events occurring over finer or coarser daily periods.

A day in a user's life is finally represented as a *multimodal bag of words*, where a word is a location sequence, concatenated with the corresponding proximity group and a timeslot, as shown in Figure 4.1. The bag of word model is amenable for probabilistic topic modeling which is introduced in the next subsection.

4.2 Modeling

In this chapter, we use LDA for two tasks:

Routine Discovery: We propose to extend the use of LDA to handle multimodal data, expecting that topics will capture joint patterns of location and proximity that help disambiguate relevant cases (e.g., discriminating between a person at work alone and in a group). Routines can be identified by observing the top words for a given topic (ranked by their probability) and also by the top days for a given topic.

Predicting Behavior: LDA is also used for the prediction of missing labels in a day (i.e., the prediction of users' joint patterns of location and proximity for certain timeslots). To achieve prediction, LDA inference is run on the test days containing missing bits. The algorithm details are presented in Figure 4.2. s_m is defined as the timeslot of a document (a day), where $m = 1...8$ are the 8 coarse-grain possibilities in a day. After finding topics within the training corpus via LDA, a distribution of topics for each test day, $\tilde{d}$, is inferred resulting in $\hat{\theta}^{(t)}_{\tilde{d}}$. The resulting topics for days $\tilde{d}$ are ranked according to $\hat{\theta}^{(t)}_{\tilde{d}}$ and the best matching topic for day $\tilde{d}$ is denoted by $z^{top}_{\tilde{d}}$, which is found according to Step 4 in Figure 4.2. The result is a single topic which is used for replacement of the missing data over the timeslot. To fill in the missing location and proximity words, we replace the missing labels with those of the top day for the mostly likely topic selected; $\tilde{d}(s_m) = d_{z^{top}_{\tilde{d}}}(s_m)$, where $d_{z^{top}_{\tilde{d}}}$ is the most probable day given $z^{top}_{\tilde{d}}$. For the predicting behavior task (whose results are given in Section 4.3.4), experiments are performed over 10 chains. Note that the procedure used for behavior prediction described here is simple and more elaborate methods to predict missing labels could be derived from the output generated by LDA.

4.3 Experiments and Results

4.3.1 Data and Model Parameters

These experiments are performed on all of the 97 individuals in the RM dataset and with days ranging from 18.07.2004 to 05.05.2005, encompassing 291 consecutive days. This subset of days was chosen since these are the days for which proximity data is mostly available. Days with entirely no reception for location were not considered since they contain no useful information for proximity either. The LDA model for joint location-proximity routine discovery used $T = 100$ topics. Heuristic methods were used to obtain T in this case, but generally speaking, a small value of T will produce coarse routines, whereas a large T will be much more specialized but can also produce redundant topics. The hyperparameters were set to $\beta = 0.01$ and $\alpha = 50/T$. These hyperparameters are chosen based on standard values used for text analysis [38].

4.3.2 Exploratory Analysis

We performed an analysis of the proximity data to study the interactions of business students compared to engineering students and staff, considering interactions for different days of the week as well as

// GOAL: Given a test day $\tilde{d}$ with missing location and proximity labels for timeslot s_m, predict a label.

// Topic discovery from the training corpus.
1) The Gibbs sampling algorithm in Figure 3.5 is performed on the training corpus to discover topics.

// LDA querying is performed to retrieve days relevant to test days.
2) Follow the Gibbs sampling procedure in Figure 3.5, replacing the topic sampling in Step 5 by the following equation, from which topic t is sampled:

$$p(z_i = t|\mathbf{z}_{\neg i}, \mathbf{w}) \propto \frac{n_t^{(w)} + \tilde{n}_{t,\neg i}^{(w)} + \beta}{\sum_{w=1}^{V} n_t^{(w)} + \tilde{n}_{t,\neg i}^{(w)} + \beta} \cdot \frac{n_{\tilde{d}\neg i}^{(t)} + \alpha}{\sum_{t=1}^{T} n_{\tilde{d}}^{(t)} + \alpha}, \tag{4.1}$$

where $\tilde{d}$ represents the test day, and $\tilde{n}_{t,\neg i}^{(w)}$ counts the observations of term w and topic t in the test days, excluding the i^{th} index [42].

3) The day/topic distribution for the test day $\tilde{d}$ is $\hat{\theta}_{\tilde{d}}^{(t)} = \frac{n_{\tilde{d}}^{(t)} + \alpha}{n_{\tilde{d}} + T\alpha}$.

// Find the best matching topic for test day $\tilde{d}$.
4) The topic z_i for which $i = \underset{j}{\operatorname{argmin}}|s_j - s_m|$ and $\hat{\theta}_{\tilde{d}}^{(z_i)} > Th$, where s_i is the timeslot of the most probable word of topic z_i, s_m is the timeslot of the missing data, $s_i, s_m = \{0, ..8\}$, and Th is a threshold, is chosen as $z_{\tilde{d}}^{top}$.

//Replace the missing data.
5) $\tilde{d}(s_m) = d_{z_{\tilde{d}}^{top}}(s_m)$, where $d_{z_{\tilde{d}}^{top}}$ is the most probable day given $z_{\tilde{d}}^{top}$.

Figure 4.2. Algorithm for predicting proximity and location timeslots.

interactions with others in the same group compared to other Bluetooth devices (not including people in the group), which could include family members, friends, strangers, laptops, or computers. The results are illustrated in Figure 4.3. The entire Reality Mining dataset was considered for these results, including 16 months of 97 users' data.

Figure 4.3 (a) and (b) illustrates the quantity of interactions between users of the Reality Mining study. Users 1-27 are the Sloan business students, and users 28-97 are the Media lab students and staff. There are two boxes marking the separation between those groups in Figure 4.3 (a) and (b). We plot the quantity of interactions between individuals in terms of (a) the number of interactions during the course of the study (without taking into account the duration of interaction) as well as (b) the total duration of interaction between these users in hours. In both plots, the amount of interaction (either considering number of interactions or total duration) was much higher between several Media Lab users, in comparison to business students. The figures have been adjusted to visualize the interaction between business students as well by assigning any interactions occurring over a threshold to the last bin of the

colorbar (200+ interactions or 150+ hours). More specifically, in Figure 4.3 (a), if there are 200 or more interactions between a pair of users, this is labeled by 200+. The threshold is chosen by rounding up the maximum number of interactions between business students. The same procedure is applied in Figure 4.3 (b) for hours of interaction. The maximum number of interactions throughout the study occurred between a pair of Media Lab users, and was approximately 585. The maximum duration of interactions occurred between a differing pair of Media Lab users, and was on the order of 690 hours over the course of 16 months. Note that these plots are not symmetric due to the inconsistencies in Bluetooth. Often times, two people will be sensed as being proximate only by one of the phones. Furthermore, there are several users without any data recordings. There are many interactions which occur frequently between individuals though not for long durations. This is especially visible between several of the Media Lab users. Also note that interactions between business students and engineering students are quite sparse. There are many Media Lab users that never interact, though most business students (with data recorded) interact, resulting in a much less sparse matrix. There was a pair of users with negative duration values, likely due to incorrect clock settings, which was removed.

Figure 4.3 (c) and (d) plots the overall means of the number of interactions and the duration of interactions in hours respectively, for Media Lab and business users over the week where 'S M T W T F S' on the x-axis corresponds to 'Sunday through Saturday'. These average values varied greatly across users. We can see in both groups of people, the interactions are very low on the weekends. The mean number of interactions is always on average higher for Media Lab students on every day of the week, though it is especially higher on Mondays, Wednesdays, and Fridays. Note however that the Media Lab population is larger and the results here are not normalized by population size. The duration of interactions for Sloan students is on average higher than Media Lab students on Thursdays, perhaps due to a course or business school event on this day.

In Figures 4.3 (e) and (f), we plot the total duration of user interactions with users in the study compared to 'non-user' Bluetooth devices (or other devices), which could include family, friends, strangers, laptops, and computers. Figure 4.3 (e) illustrates the total interaction times of Sloan users whereas 4.3 (f) is for the Media Lab users. In 4.3 (e) and 4.3 (f) we can see there are a few people in both groups who have heavy interactions within the group. Also, many of the users have more interaction with people in the group than with 'other devices'. Many of the Media Lab users have heavy interactions with 'other devices', likely due to the fact that they spend hours in front of their laptops and computers daily.

4.3.3 Joint Location-Proximity Routines

The fusion of proximity and location data enables the discovery of more detailed patterns regarding this group of MIT users' daily lives compared to single modalities. After LDA learning, there is a chance that two topics could be similar to each other, as LDA does not guarantee that topics be distinct from each other. The fact that LDA-learned topics are often similar to each other has also been observed in the text domain. A short summary of the learned routines on the entire corpus is presented below, and a summary is visualized in Figure 4.4.

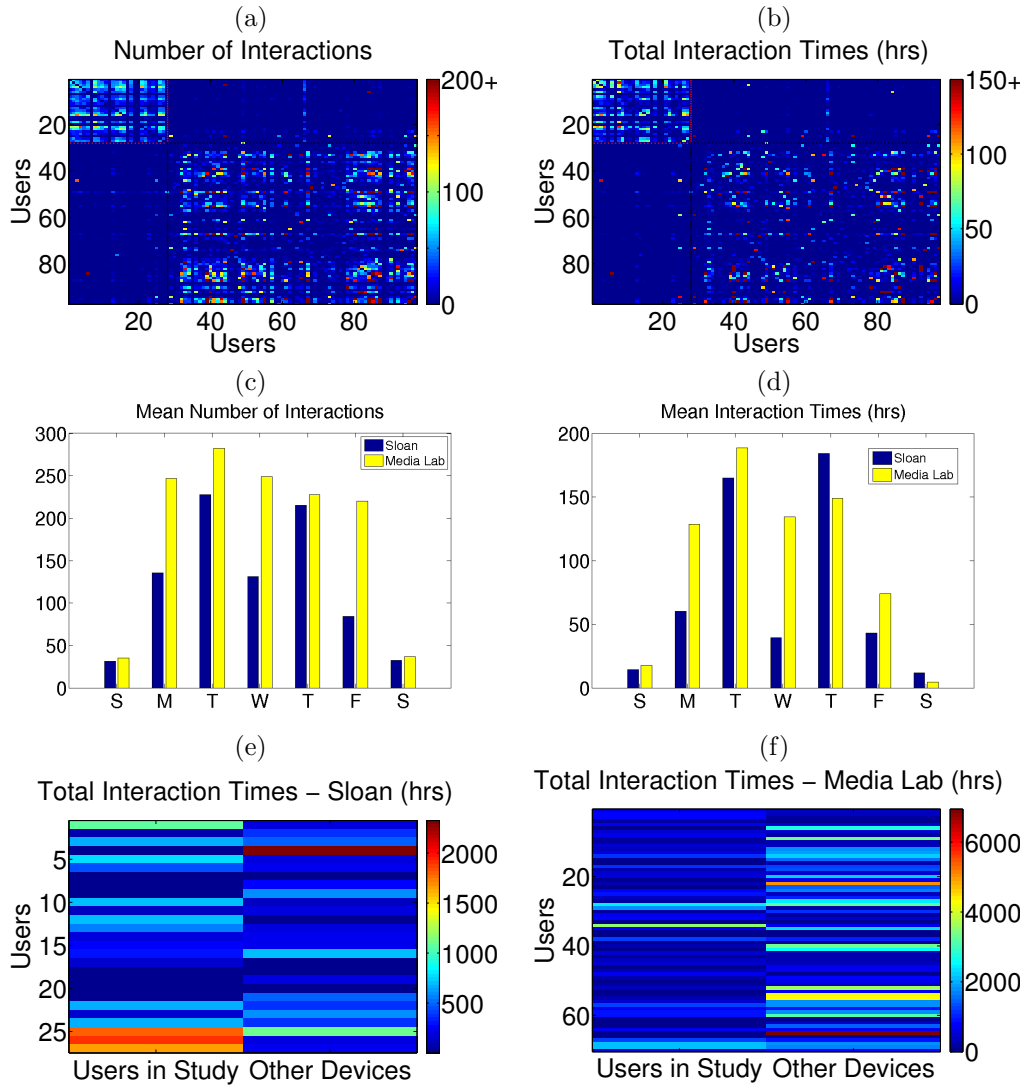


Figure 4.3. Interaction patterns of MIT business students compared to engineering students and staff. Figures (a) and (b) visualize the pairwise user interactions in terms of (a) the number of interactions and (b) the total duration of interactions (hours). Business students (1-27) and Media Lab users (28-97) are highlighted by boxes. There are many interactions between engineering students which do not occur over long durations. The average quantity of interactions over all Sloan business students versus all Media lab students and staff is computed over the days of the week 'S M T W T F S' in terms of (c) the number of interactions and (d) the total duration of interactions (hours). On average, Media Lab users have more interactions, though on Thursdays business students interact for longer durations, perhaps due to a course on this day. They also interact less on Mondays, Wednesdays and Fridays. On average, there is little interaction on weekends in all cases. The total interaction times (hours) of users with other Reality Mining users in comparison to all other Bluetooth devices are shown in (e) for Sloan students and (f) for Media Lab users.

Topic 11 Top 3 words	$p(w\|z)$
W W W 11 am-2 pm Small Group	0.502
W W W 2 pm-5 pm Small Group	0.427
H H W 9 am-11 am Dyad	0.015

Topic 28 Top 3 words	$p(w\|z)$
O O O 9 pm-12 pm Alone	0.634
O O O 7 pm-9 pm Alone	0.358
W O W 11 am-2 pm Dyad	0.001

Topic 41 Top 3 words	$p(w\|z)$
W W W 2 pm-5 pm Large Group	0.610
W W W 5 pm-7 pm Large Group	0.275
W W H 5 pm-7 pm Large Group	0.034

Topic 46 Top 3 words	$p(w\|z)$
W W W 7 pm-9 pm Small Group	0.415
W W W 9 pm-12 pm Small Group	0.219
W W W 5 pm-7 pm Small Group	0.132

Topic 53 Top 3 words	$p(w\|z)$
H H H 0-7 am Alone	0.757
H H H 7 am-9 am Alone	0.215
H H W 9 am-11 am Alone	0.015

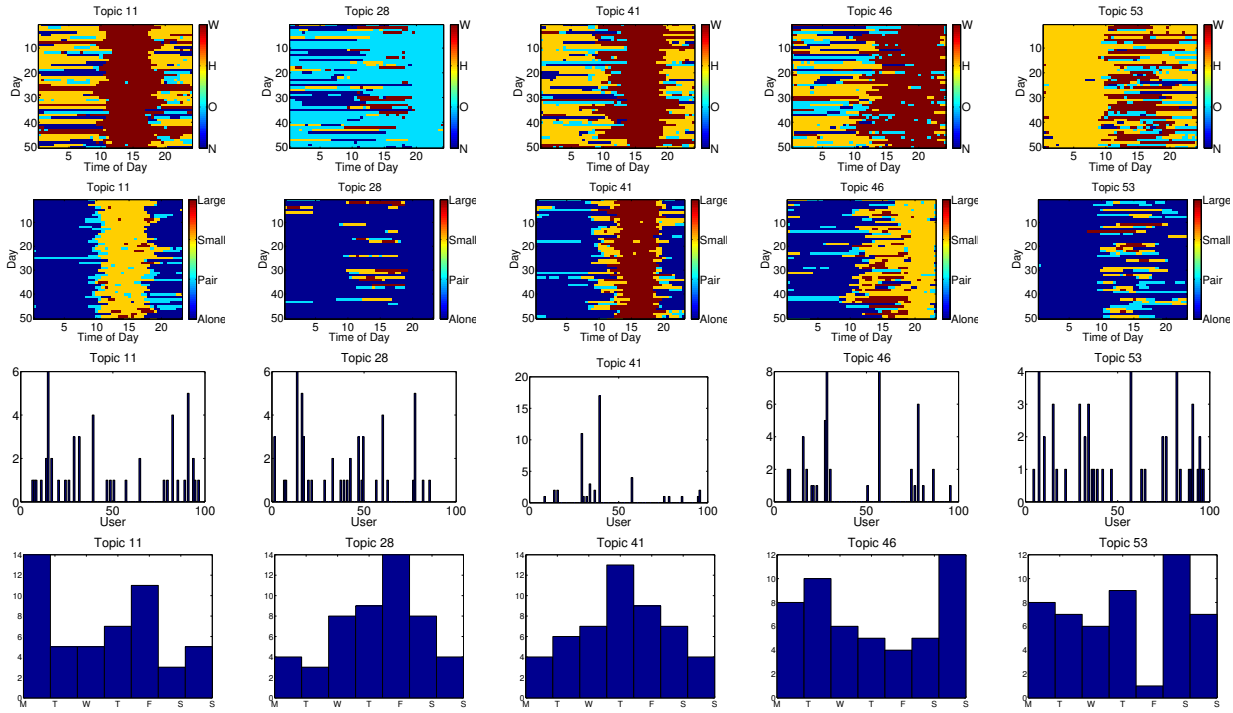


Figure 4.4. Selected LDA results. The first row of tables correspond to the most probable words given a topic. Ranked days (i.e., documents) for selected topics by $p(d|z_i)$, showing (second row) the top 50 days' location data and (third row) the corresponding proximity data for a given topic. (fourth row) Histograms of the users whose days ranked in the top 50 for topic z_i. (last row) Histograms of the days of the week (M T W T F S S = Monday to Sunday) that ranked in the top 50 for topic z_i. Note the colorbars for the location figures indicating the W, H, O, and N locations, and for the proximity figures indicating a large group, small group, pair, or alone.

-*Home routines and proximity:* Most of the home routines discovered occurred for users alone (i.e., not in proximity with anyone from the group). Only 2 out of the 20 topics related to discovered home routines dominated for a pair of users in proximity. No home routines occurred for small or large groups in proximity, which suggests that people did not socialize within the population at home.

-*Work routines and proximity:* Most of the routines discovered with proximity interactions occurred at work locations. There are 17 topics corresponding to work routines, and 13 of them occur with proximity patterns. Routines at work were discovered for all four proximity groups (users alone, in dyads, small, and large groups), which indicates that all these types of interactions occur frequently.

-*Morning routines and proximity:* Only 3 out of 100 topics had a proximity interaction in the morning (before 10 am), and all 3 of these routines occur for pairs of users and never for groups. People interacting in the morning seem to be relatively sparse for this population.

-*Day time routines and proximity:* Approximately 20 topics characterize user interactions throughout the day (10 am-7 pm). The interactions include pairs of users, as well as small and large groups.
-*Evening routines and proximity:* 7 topics characterize group interactions in the evenings (7 pm-midnight). These occur for pairs of users, and small as well as large groups.

A selection of topics illustrating the types of joint routines discovered are visualized in Figure 4.4. We have illustrated results for selected topics, $z_i = 11, 28, 41, 46, 53$, for the 50 most probable days given those topics. The three most probable words given the topics are shown in the tables in the first row. We plot the results in terms of users' locations (second row), proximity (third row), user statistics (fourth row), and day of week statistics (fifth row). The figures illustrating the users' locations and proximity data show the time of the day as the x-axis, and each row is a day of the life of a user plot in terms of their location (H is home, W is work, O is out, and N is no reception) as well as in terms of their interactions where ('large' corresponds to a large group, and 'small' to a small group). Furthermore, a histogram for the users whose days ranked in the top 50 days is shown in the fourth row, the x-axis indicating anonymous user id and the y-axis the number of days. The fifth row illustrates a histogram of the days of the week (M T W T F S S = Monday to Sunday) of the 50 most probable days given each topic. A summary of the routines discovered plotted in Figure 4.4 is:
-**Topic 11**: The user is at work during the day (dominantly 11 am-5 pm as seen from the 3 top words given topic 11) while in proximity with a small group of 3-5 people. Several users have days with high probability of topic 11. This work routine dominates on Mondays.
-**Topic 28**: The user is out in the evenings (7 pm-12 pm) alone. This routine occurs most frequently on Fridays for several users in the study.
-**Topic 41**: The user is at work from 2 pm-7 pm in a large group. This occurs dominantly for a handful of users, predominantly on Thursdays. Note that most of these users correspond to Sloan business school students, displaying their common Thursday afternoon work routine.
-**Topic 46**: The user is at work in the evening (from 5 pm-midnight) in a small group. This work routine dominates on Sundays and occurs often for a few users.
-**Topic 53**: The user is at home alone in the mornings (from midnight until 11 am). This topic rarely occurs on Fridays.

Overall, we find topics expressing the dominant socio-geographic routines in the data. We now investigate predicting missing multimodal data in order to objectively evaluate the methodology.

4.3.4 Behavior Prediction

We now show how it is possible to use LDA in order to predict unobserved location and proximity data for a timeslot of a user's day. For experiments, we decided to distinguish between people based on the entropy of their routines under the hypothesis that prediction of location and proximity will be more or less difficult depending on the variability of each person's habits. User entropy is computed on the distribution of topics given users, $p(z|u) = \sum_d p(z|d,u)p(d|u)$, where u is the user variable, and we assume $p(d|u) = \frac{1}{|D_u|}$, D_u is the set of users recorded for user u, and $|D_u|$ is the set cardinality. The

topics z correspond to the joint location-proximity routines found in Section 4.3.3. All of the users in the dataset are ranked according to their entropy. After this, we set two thresholds for high and low entropy which gave 10 users in each case. We randomly picked 5 people for each class (high and low entropy).

For each of the 10 selected users, 20 days of their life were randomly selected from days with at least one proximity interaction (i.e. days that contained at least one non-empty word over the entire day). This set of days was used to form the test set, from which we systematically remove words to generate data with missing sequences to predict. For each day, the words of a given coarse-grain timeslot were removed to form a day for which the method has to predict the missing sequence, thus generating 8 days, each with one timeslot's words missing. The resulting dataset for which we predict missing sequences contains 10 users, each with 160 days = 1600 days for testing. Thus, for each user there are 160 days for testing, and each coarse-grain timeslot contains 200 days for testing.

For each day, there is one timeslot with missing location and proximity labels. For evaluation, we compute two types of error.

- The *location error* is the number of incorrectly predicted location labels divided by the total number of labels to be predicted in the given coarse-grain timeslot. For instance, days with timeslot 1 missing have 14 location labels to be predicted since it occurs from 0-7 am.
- The *proximity error* is the average number of people wrongly predicted for each word in a given timeslot. More specifically, if the predicted group (alone, dyad, small group, large group) is correct then there is no error. If the predicted group is incorrect, then we predict the minimum number of possible people in the group (alone=1, dyad=2, small group=3, large group=5) and compute the difference with the actual number of people in proximity. For example, if there are 10 people in proximity and we predict a small group, then we assume 3 people are in proximity. If this incorrect prediction occurs over the 14 half-hour words in timeslot 1 (midnight-7 am), then the average proximity error is 7.
- Finally, the results for location and proximity error are averaged over 10 randomly initialized chains of the Gibbs sampling procedure described in Figure 3.5.

The location and proximity errors are computed over users and timeslots and displayed in Figures 4.5 and 4.6. We present the average errors as a function of the user for location in Figure 4.5 (a) and for proximity in Figure 4.5 (b). Users 1-5 (in blue) have low entropy and 6-10 (in red) have high entropy. Interestingly, low-entropy users have lower error in the prediction of location labels than high-entropy users. For low entropy users, the error can be as low as 0.32 which nevertheless indicates that the task is difficult. We also include errorbars corresponding to the standard deviation over the 10 randomly initialized chains. High entropy users are significantly more difficult to predict. In Figure 4.5 (b) we plot the proximity error. In the best (resp. worst) case, the predicted number of people in proximity is incorrect by, on average, 0.16 (resp. 1.2) people. In this case, low entropy users do not necessarily have lower prediction errors in proximity than high entropy users.

In Figures 4.6 (a) and (b), we plot the average errors as a function of coarse-grain timeslot for both

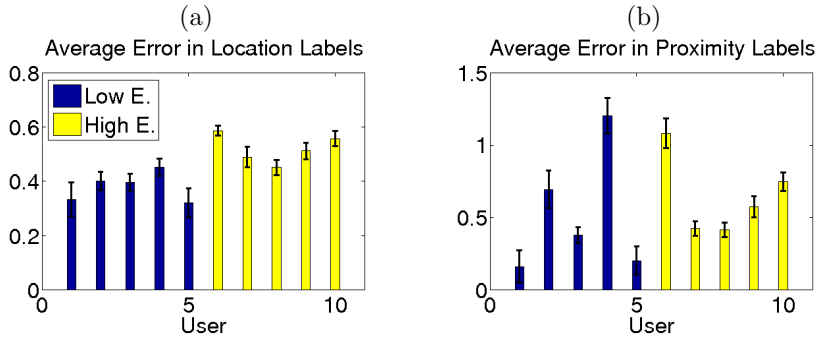


Figure 4.5. (a) Average location prediction error as a function of users, where low entropy users are labeled 'Low E' and high entropy users 'High E'. (b) Average proximity prediction error as a function of users. Location label for prediction is consistently lower for low entropy users. However, for proximity errors are not necessarily lower for low entropy users.

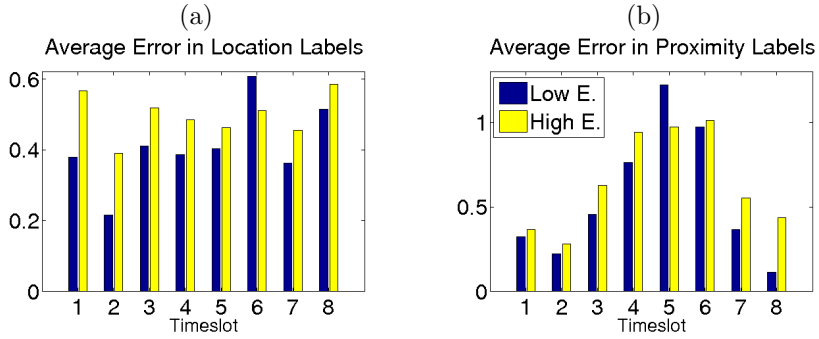


Figure 4.6. Average error in (a) location prediction, and (b) proximity prediction, as a function of timeslot for low and high entropy users. High entropy users consistently have higher location label errors for prediction over all times of the day, though the error is highest between 5-7 pm (timeslot 6) which corresponds to typical commuting times. The highest errors in proximity label prediction occur from 9 am-7 pm, corresponding to work times where most interactions occur.

high and low entropy users for location (Figure 4.6 (a)) and proximity (Figure 4.6 (b)). We can see in Figure 4.6 (a) that for almost every timeslot (with the exception of timeslot 6), high entropy users are harder to predict (have higher errors) than low entropy users. Timeslot 6 (5-7 pm, which corresponds to typical commuting times) is overall the most difficult to predict. Also, for timeslots 1 and 2 (midnight to 9 am), low entropy users correspond to much better performance than high entropy users. Regarding Figure 4.6 (b), the error in proximity prediction as a function of timeslot is again not highly correlated with the entropy of a user. The prediction in proximity has the highest error in timeslot 5, corresponding to 2-5 pm, and the lowest error in the mornings and late evenings, which is not surprising. In the worst case, the proximity error in any given timeslot is less than 1.25 people on average.

In Figure 4.7, we compare the performance of our topic model (TM) method to several methods. Figure 4.7 (a) illustrates the overall average location error for the TM approach in comparison to a nearest neighbor approach called previous day (PD), which uses knowledge about the specific date of the test day, and replaces the missing data with that of the previous day. Note that the date is a very strong

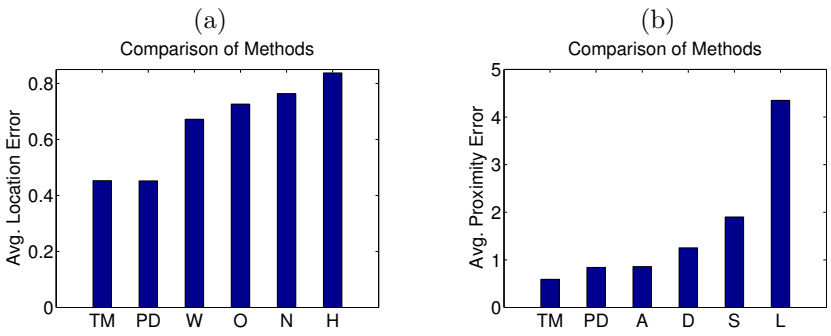


Figure 4.7. A comparison of our topic model (TM) approach to various other methods for overall location and proximity errors. PD is the nearest neighbor approach of replacing data with the previous days'. (a) For the overall average location error W represents the error obtained if all missing data is replaced by work, O by other, N by no reception, and H by home. (b) For the overall average proximity error A represents the error obtained if the missing data is replaced by alone, D by dyad, S by small group, and L by large group. The TM approach predicts missing location data as well as the PD approach, however, our approach outperforms the PD method for predicting missing proximity data. The TM also outperforms all the other methods (both in terms of location and proximity missing data prediction) significantly.

contextual cue about human routines that is not currently used in our method TM. The approach labeled W is the case where all missing data is replaced by the 'work' location. Similarly O is the case where all missing data is replaced by 'out', N by 'no reception', and H by 'home'. Figure 4.7 (b) illustrates the overall average proximity error for the TM approach in comparison to the PD approach, in addition to the approaches labeled A, D, S, and L, corresponding to replacing the missing proximity data with the labels 'alone', 'dyad', 'small group', and 'large group', respectively. We can see the TM and PD approaches perform similarly in terms of location data prediction, however, TM outperforms PD for missing proximity data prediction. The TM approach also outperforms all the other methods.

Given the simplicity of the PD method, we look deeper into the TM and PD performance for various types of users. Figure 4.8 (a) illustrates the average location prediction error for high and low entropy users. We see that for location prediction, the PD method performs better for low entropy users. This is understandable since low entropy users have very 'routine' lifestyles and simply replacing the missing data with that of the previous day results in good performance. However, for high entropy users, our TM method, which captures specific patterns of transitions (e.g., H to W) works better. Given these complementary features, for future work, we plan to investigate a method that integrates both concepts. Figure 4.8 (b) illustrates the average proximity prediction error for high and low entropy users. The results show that our TM approach outperforms the PD approach both for low and high entropy users.

4.4 Classifying Daily Routines

In this section we move from an unsupervised approach to a supervised one, based on support vector machines. Supervised machine learning models require training data with labels and are thus only suitable for tasks in which some prior known knowledge of the features is available.

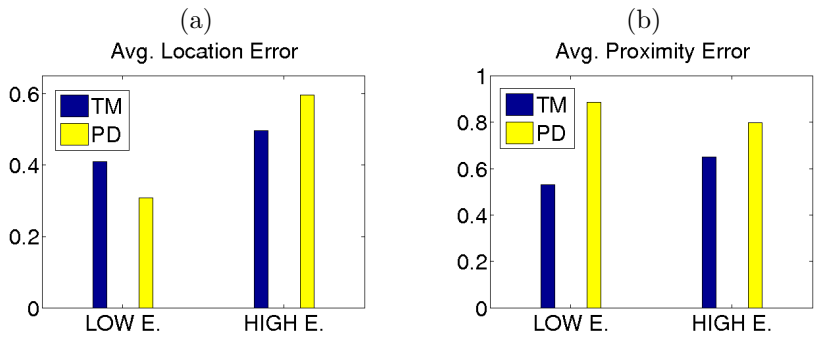


Figure 4.8. Comparison of our topic model (TM) approach with the previous day (PD) approach in terms of user types. (a) Average location error for low entropy users and high entropy users for the TM vs. PD approach. The PD approach performs better for location data prediction for low entropy users, however, our TM approach performs better for high entropy users. (b) Average proximity error for low and high entropy users for the TM vs. PD approach. In both cases our TM approach outperforms the PD approach for proximity data prediction.

We classify people's daily routines defined by day type, and by group affiliation type. We propose and compare single- and multi-modal routine representations at multiple time scales, each capable of highlighting different features from the data, to determine which best characterized the underlying structure of the daily routines. We show that the integration of location and social context and the use of multiple time-scales is effective, producing accuracies of over 80% for the two daily routine classification tasks investigated, with significant performance differences with respect to the single-modal cues. We address two classification tasks for daily routines: weekday vs. weekend routines, and engineering student-like vs. business student-like routines. In both cases, the input data is one day of location and/or proximity information.

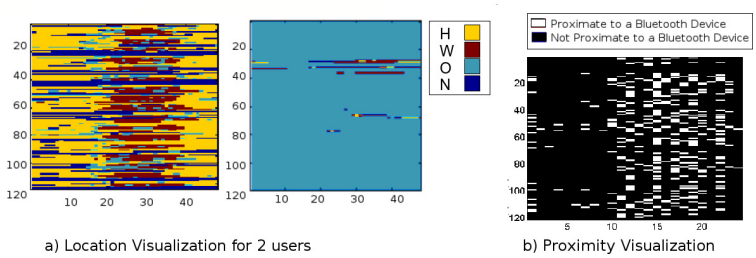
4.4.1 Data Representation

The goal is to represent a day using location and proximity information that is discriminant to daily pattern classification. A day can be represented at multiple time scales, and people's routines usually follow block-type schedules. As done in the previous sections we quantify location and proximity information at two levels (one fine-grain at 30 minutes and one coarse-grain at 3-4 hours). These two time scales provide a simple model of time management that is appropriate to characterize many people's lives. We use various location-driven and proximity-driven representations presented next.

Location Representation

L_a *Fine-Grain Location.* For the fine-grain location representation previously described in Section 3.2.1, a day is divided into 30 minute non-overlapping time intervals, resulting in 48 location labels (W, H, O, or N) per day. For classification purposes, this 48 element vector was transformed to binary format.

L_b *Bag of Location Transitions.* This representation, previously described in Section 3.2.2, was built



a) Location Visualization for 2 users b) Proximity Visualization

Figure 4.9. (a) Visualization of location patterns using the fine-grain location representation, L_a, for 2 users over 121 days. Each row in the graph represents a day in the life of the user. The user on the left has a rich set of routines visible in the location patterns, whereas the user on the right is mostly incomplete due to lack of celltower labels. (b) The proximity representation, P_b, is visualized for a user. Only proximity with users in the group are considered. For this user, most proximity activity occurs later in the day for most days.

from the fine-grain location representation considering 8 coarse-grain timeslots in a day. The bag of location transitions is the histogram of the 4 component location sequences in the day.

L_c *Coarse-Grain Location.* For this representation, visualized in Figure 4.10 (a), a day is also divided into 8 coarse timeslots. For each timeslot, there is a binary element representing the four location labels (H, W, O, N). If one of these labels was present within the given timeslot, it is counted as one, if this location was not present, it is counted as zero. This is a simplification of the bag of location transitions, in which the dimensionality was reduced to be comparable to some of the proximity representations described in the next subsection.

L_d *Two-Feature Location.* This representation is the simplest, in which the number of 30-minute H and W labels are counted, without taking into account when exactly they occur in a day.

Proximity Representation

P_a *UserID Proximity.* The userID proximity representation is illustrated in Figure 4.10 (b). There are 31 binary components for a given day, reflecting the 30 people considered for classification (see Section 4.4.3), and the last component for the case when no one is in proximity. If the person was in proximity with one of the 30 individuals, the value for that component will be one; for days when the person is not in proximity with anyone, the last component will be one. Thus, we only consider proximity within the set of 30 people. Obviously, we do not consider a person to be in proximity with oneself.

P_b *Coarse-Grain Proximity.* The coarse-grain proximity representation, visualized in Figure 4.9 (b), contains again the same 8 timeslots in a day. In this description of proximity, the identities of people are disregarded and only the quantity of proximate people for a timeslot is considered. In the first timeslot, the first bin element is one if 1 to 2 people are in proximity, the second bin if 3 to 4 people are, and the last if 5 or more people are in proximity. The resulting representation contains 8 timeslots, each with 3 elements. This idea of binary quantization is repeated over the 8 timeslots, resulting in a quantification of interaction within the total set of people over different times in the day.

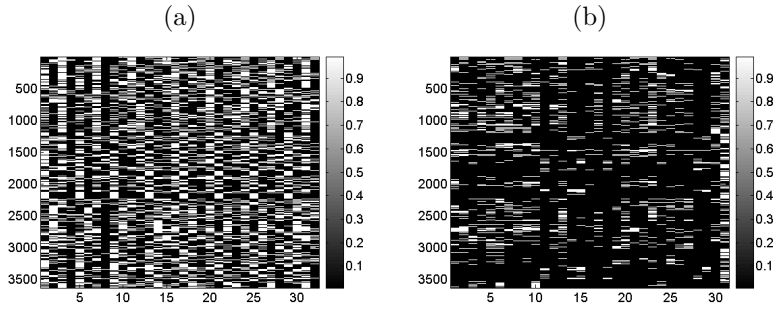


Figure 4.10. (a) Coarse-grain location representation, L_c, visualized over all days and users. (b) UserID proximity, P_a, displayed over all users and days.

P_c *One-Feature Proximity.* This is the simplest representation for proximity. We count the number of proximate people for a person within a day, and use this value.

Combined Representation

For the combined representation, we concatenate one of the location representations with one of the proximity representations. We only consider cases with comparable location and proximity dimensionality. Feature extraction techniques (e.g., PCA) could have been applied on the joint representations but were not explored here.

4.4.2 Classifier

The classification was performed using a support vector machine (SVM) with a Gaussian kernel. For both daily routine classification tasks (days as weekends or weekdays, or days as belonging to business students or engineering students), the training strategy was leave-one-user-out, specifically testing on all the days for one unseen person while training on the data for all other people (note: proximity features are by definition relational, involving pairs of people); we tested on each of the people and averaged the results. We optimized the kernel parameter on one data split for a randomly chosen person.

4.4.3 Dataset

From the RM dataset, we considered 30 people and 121 consecutive days, resulting in approximately 3600 data points. Our choice was guided by the goal of analyzing people and days for which data was reasonably available. The exact dates in the experiment were 26.09.2004 to 24.12.2004. The people selected had the most number of days with at least one W or H label. We removed days which were entirely N (no reception) labels, which resulted in approximately 2800 data points. To select the interval of 121 days, we found the time interval with the most number of useful days (i.e., days with W, H, or O labels) over all 30 people. The resulting subset amounts to over 87000 hours.

For the student-type daily routine classification task, a subset of 23 of these 30 people were considered

based on their student type labels. There were 6 business school students, and 17 engineering students. The engineering students covered a broader scope, including both undergraduate and graduate levels.

4.4.4 Weekday/Weekend Routine Classification

The weekend/weekday classification results are presented in Table 4.1 and reveal the difficulty of the task solely based on location or proximity information. In each table, the classification accuracy averaged over all people is presented first, and the average accuracy for each class is presented later. Generally, weekdays are more easily identified with location as input, and weekends are characterized better by proximity data. We can understand this by identifying weekdays with WORK cell towers, and weekends by not being in proximity with colleagues. However, in this dataset, students appearing to be in W locations on weekends complicate the classification task, resulting in at best 44.1% weekend classification accuracy by the bag of location transitions (L_b), which performs overall better than the others, also having the highest dimensionality. The coarse-grain approach L_c (fused bag of location words) performs slightly worse for weekends with a significantly smaller dimension. The fine-grain location representation, L_a, performs the worst for WE, the best for WD, and slightly better than the two-feature location case. All methods perform better than a 'naive' guess that assumes all days are weekdays, which results in $5/7 = 71.4\%$ accuracy.

Proximity information alone is useful in characterizing weekends, but does not perform as well as location data for identifying weekdays. There are many weekdays with little group interaction, resulting in higher confusion with weekdays. The userID proximity and one-feature cases (P_a and P_c) reveal about 2% difference between their weekend and weekday performances, overall resulting in the highest performance of approximately 74%.

The lower panel in Table 4.1 shows the improvement in classification with the combination of proximity and location data. Note that in all cases the overall performance of the joint representations improved over that of the singleton case. We achieved over 80% accuracy with the combined representation (L_c, P_a) trading-off 2-3% weekday accuracy for improved weekend classification. In Figure 4.11, we visualize the days for which the proximity-alone data (columns 33-56) was misclassified, however when we added the location data (columns 1-32), the resulting 56-component vectors were correctly classified. In both figures, the first 32 columns visualize the location representation L_c and the last 24 columns illustrate the proximity representation P_b, so each row displays a day of the combination (L_c, P_b). Figure 4.11 (a) are weekends which performed incorrectly for proximity-alone data due to the abundance of proximity interactions, which are not typical of weekends. In contrast, Figure 4.11 (b) shows weekdays which were mistaken for weekends due to the sparsity in interactions, not typical of weekdays. The addition of the location information in both cases resulted in correct classification, thus illustrating cases for which the combination of information improved classification performance.

The performance difference between the best location only method (L_b) and the best combined method (L_c, P_a) is statistically significant at the 1% level. The same is true for the performance difference between the best proximity only method (P_c) and the best combined method (L_c, P_a).

Table 4.1. Weekend (WE) and Weekday (WD) daily routine classification accuracy. The top table shows the difficulty in determining weekends based on location information alone. Proximity data is more deterministic of weekend routines. Classification obtained by combining location and proximity results in the best performance.

Location Accuracy (%)				Proximity Accuracy (%)			
Method	**Overall**	**WE**	**WD**	**Method**	**Overall**	**WE**	**WD**
L_a	74.2	19.3	95.3	P_a	74.3	70.7	75.8
L_b	76.8	44.1	89.1	P_b	72	54.2	78.7
L_c	76	36.6	90.8	P_c	74.6	67.9	77.1
L_d	75.7	30	93.1				

Combined Accuracy (%)			
Method	**Overall**	**WE**	**WD**
(L_d,P_c)	76.9	47.35	88.1
(L_c,P_a)	80.3	65.8	85.8
(L_c,P_b)	79	53.4	89.3
(L_a,P_a)	76.5	60.2	82.8

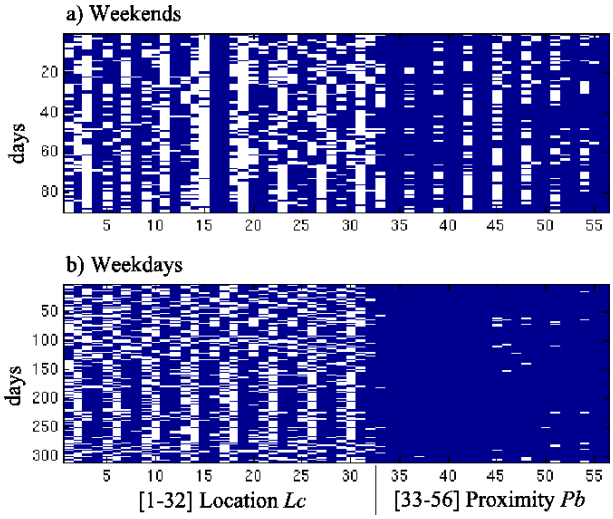


Figure 4.11. Advantages of the joint location-proximity representation (L_c, P_b). Visualization of (a) weekends and (b) weekdays for which the proximity-only data was misclassified, but for which the location-only data and the combined proximity-location data were correctly classified. The sparsity of the weekday proximity-only data (columns 33-56 in (b)), resulted in incorrect classification since sparsity in interaction is typical weekend behavior. However, when we added the location information, the resulting combined representation was correctly classified. The opposite phenomena can be observed in plot (a), for which weekends have abundant proximity data, typical of weekday behavior.

Table 4.2. Engineering (Eng) vs. Business (Bus) student daily routine classification results. Proximity within the specific group is most representative of student type, especially when student identity is retained. The joint location and proximity data improves classification performance for the (L_c, P_a) combination. However, the other combinations generally perform as well as the singleton cases.

Location Accuracy (%)				Proximity Accuracy (%)			
Method	**Overall**	**Eng**	**Bus**	**Method**	**Overall**	**Eng**	**Bus**
L_a	66.8	90.4	0	P_a	89.1	98.9	61.2
L_b	74.54	94.3	19	P_b	78.1	96	28.1
L_c	74.5	94.8	17.1	P_c	50.2	95.3	0
L_d	74.8	99.6	4.5				

Combined Accuracy (%)			
Method	**Overall**	**Eng**	**Bus**
(L_d,P_c)	73.3	97.6	4.5
(L_c,P_a)	89.6	99	62.9
(L_c,P_b)	78.76	93.4	37.4
(L_a,P_a)	84.5	95	54.7

4.4.5 Business/Engineering Student Routine Classification

Effectively classifying daily routines as belonging to business students or engineering students based on proximity-only observations was representation-dependent. The results are presented in Table 4.2. Proximity representation P_c, the one-feature case, was inadequate in differentiating between student types, suggesting that the overall quantity of proximity within each group is on average the same. If the business students had much more proximity within the total set of people, or vice versa, we could expect the one-feature case to produce higher accuracy. The coarse-grain proximity representation P_b improved the accuracy of business student classification, however, the userID proximity representation proved to be the best, with almost 99% accuracy in engineering student classification and 61% for business students. The knowledge of identity from proximity is the key for discriminating student disciplines.

Location knowledge was inadequate in student type determination for the most part. This is likely due to the simplified representation used where the 32 000 cell tower IDs have been reduced to four location classes. It is expected that a representation more precisely identifying the geographic location of a student would perform better. However, the representation used here is useful in understanding whether student types differ in the amount of time spent at school, home, or out and about. The two-feature location case, L_d, having low accuracy, indicates that the amount of time spent at school and home is not indicative of student type. The most effective characteristics in differentiating, which can be observed in Table 4.2 by the highest performance with the bag of location transitions representation, might be patterns of "going to work" in a timeslot, or "coming home" in a timeslot, or other similar routines which are captured by this representation.

The performance difference between the best location only method (L_d) and the best combined method (L_c, P_a) is statistically significant at the 1% level. The performance difference between the best proximity only method (P_a) and the best combined method (L_c, P_a) is not statistically significant.

4.5 Conclusion

We have proposed a probabilistic methodology that successfully discovers recurrent patterns in people's lives from multimodal data, and that can use the discovered routines for data prediction, estimating location and proximity data of users with varying entropy. Essentially, the method mines the most dominantly occurring human routines (topics) from a huge corpus of real-life human mobile data to determine recurring human patterns involving time of the day, semantic location and proximity based interaction type. Our method also uses these rich human location-interaction topics to predict missing data, which in real life occurs very frequently with mobile phone data, and can also be seen as a method to objectively verify the validity of the routines discovered. By computing the entropy of individuals based on their jointly modeled locations and interactions, we were able to observe that our method predicts missing multimodal data over several hours for users with both low and highly varying lifestyles.

Finally, we also investigated a supervised approach to classify the daily life routines of approximately 87 000 hours of data. We achieve over 80% accuracy in identifying whether a given day more closely resembles a weekend or weekday. This is not an easy task as students spend many weekends in work locations and have many weekdays with few group interactions. We showed that the integration of location and proximity data performed significantly better than the single observation sources, and that using representations that consider multiple time scales was beneficial. We further succeeded in identifying whether a user is an engineering or business student with over 89% accuracy based on a single day pattern of activity. The identity of individuals, measured by proximity, was key in this case, which confirms that social context is very helpful to identify people's routines.

4.6 Limitations

The limitations of this work include the heuristic approach to selecting the number of topics and the model parameters. Also, the timescales used are hand-picked over several hour intervals. In future work, the methodology for data prediction could be further optimized to use the topics in a more sophisticated manner, and to include prediction on varying timescales, such as full days of missing data. It would also be very useful to take advantage of the other, often available data modalities of mobile sensor data for data prediction. For instance, one could predict a user's location given the time of day and their interactions, the day of the week, or even using their phone call and SMS data. The Bluetooth proximity data is potentially a very rich source if one considers proximity to all other devices including laptops, computers, and anonymous cell phones. This data in itself could be used to determine the semantic labels of an individual, such as if the user is at home (in proximity with their home computer), at work (in proximity with their work computer), or out (in proximity with strangers). In a different line of work, we would like to enrich the location vocabulary by refining the "other" category. This in principle could be done from the Reality Mining dataset, but handling sparse human annotation of places is in itself a research problem.

Chapter 5

Modeling Varying Length Routines

Our research in activity modeling has led to the need for the development of mathematical models that discover patterns over long and varying durations. In this chapter, we address the problem of modeling varying time duration sequences for large-scale human routine discovery from cellphone sensor data using two differing approaches. Our objective is to handle large sequence lengths based on principled procedures to deal with potentially large routine-vocabulary sizes, and be applied to simple, generic initial vocabularies to discover meaningful location-routines.

There are several difficulties to modeling human activities, as discussed in previous chapters, including various types of "noise", lack of ground truth, complexity due to the size of the data, and the various types of phone users. The fundamental issue we focus on here is that we often do not know (or cannot pre-specify) the basic units of time for the activities in question. We do know that human routines have multiple timescales, however the effective modeling of many unknown time-durations is an open problem. Previous works always assume a fixed and predefined unit of time, limiting the timescale of the routines discovered.

Previously for routine discovery we used a bag of words approach. This has also been used in previous works for activity discovery on video data [72] as well as sensor data obtained by wearable devices [48]. The advantages of a bag approach are the robustness to noise, and the compact representation. The disadvantage is that the words are often not simple and the time duration must be predefined by hand [48] or a supervised method requiring a training phase is necessary [72]. Furthermore, *previously unknown* timescales, whether single or multiple, are not considered. Here we consider two approaches to model and discover activities over varying previously undefined durations and show discovery results in terms of location routines. We also perform some experiments validating the effectiveness of the techniques, for example in terms of the model parameter growth rate and the type of sequences discovered. Some of the material presented here was originally published in [33].

The work presented in Section 5.2 was originally published in modified form in [33]. The work in Section 5.3 has not been published yet.

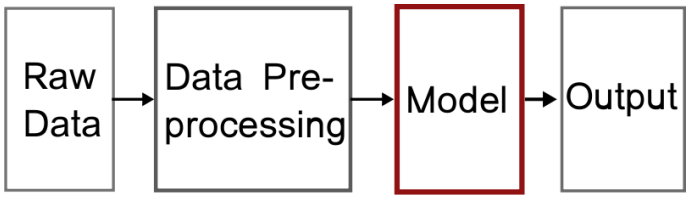


Figure 5.1. Overview of a simplified activity recognition system. We apply two approaches for the discovery of activities over varying durations. The first is the Multi-Level Topic Model (MLTM) where the idea is to use the output for preprocessing the data over multiple iterations, each time generating activities over longer durations. The second approach is the Pairwise-Distance Topic Model, where we formulate a new model to address the problem.

5.1 Overview

We consider two approaches to this problem. If we consider a simplified activity recognition system consisting of the components in Figure 5.1, raw data is preprocessed into a format which can be processed by the model. So far in this thesis the model in question is mainly LDA. The two approaches we take are as follows:

1. In the first approach, we manipulate the preprocessing unit based on the output of the model in order to achieve the goal of discovering routines over varying time durations (or varying length sequence discovery). We call this approach the Multi-Level Topic Model and present the details in Section 5.2.
2. In the second approach, we redefine the model itself to discover sequences of arbitrary length, where the length can be very large. This approach is called the Pairwise-Distance Topic Model and is described in Section 5.3

We define an n-gram to be a series of n consecutive words. We define a word, in terms of location data, to be a location label and a timestamp. More precise specifications follow, however, an n-gram in terms of activities can be interpreted as an activity occurring over a duration of $n/2$ hours if we assume words taking 30 minute intervals, as we do in this chapter. More generally, an n-gram is an activity over $n * x$ units of time where words capture data over x units of time intervals.

5.2 The Multi-Level Topic Model

In this section we propose an approach to construct vocabularies of increasingly large n-grams, built in a multi-level fashion. The number of consecutive words (or size n in the n-gram) increases with each level of the topic model. Our approach is built on LDA. The maximum number of levels determines the maximum n-gram size considered, and large n-gram sizes can easily be considered with this method. At a given level n, the vocabulary consists of only n-grams, though this could easily be modified to consider all n-grams of size n less than or equal to the level. The output of the topic model at level n is used to construct the vocabulary for level $n + 1$. There are several advantages to this technique. Firstly, the vocabulary growth is controlled and does not explode since we only consider a subset of

n-grams for vocabulary construction at each level, as opposed to the naive n-gram approach which grows exponentially with the size of the vocabulary. Secondly, not only do we consider the most frequently occurring n-grams, but we can capture frequently co-occurring n-grams, which is critical in large n-gram discovery. For example, if 2 words co-occur a lot, they are more likely to form a bigram than if each occurs very frequently, but not together. In the third place, the initial vocabulary can be very simple, not requiring much human intervention in its design. Finally, each individual level results in unique patterns with routines found for the particular vocabulary at that level. The overall contribution is an approach to identify varying length human location routines and the successful application of this idea to 10118 days and 242800 hours of location data.

5.2.1 Methodology

We consider the 97 subjects' location data in the Reality Mining dataset. We form very simple location words, capturing location and time information for a user, and use these as input to our multi-level topic model, which finds sequences of varying lengths as topic outputs in an unsupervised way. The components of our methodology are described in detail next.

Location Sequences as Words

We represent a day in the life of a mobile phone user's location as a bag of location sequences. In the simplest case (at the first level) a unigram $u_1 = (t, l)$, where t is a half-hour time interval in the day, $t \in T = \{1, 2, 3, ..., 48\}$, and l is a location label, $l \in L = \{`H', `N', `O', `W'\}$, as in Chapter 3 and 4. A bag of unigrams consists of the unigrams in the day of a user, of which there are 48. The vocabulary at level 1, V_1, consists of the set of all possible unigrams in the corpus C. The vocabularies for additional levels and n-grams for $n > 1$ are described in the following section. The bag representation we use is simple and does not require much handcraft as in previous chapters (which were essentially trigrams over coarser grain time intervals).

A Multi-level LDA

We propose a multi-level approach, as illustrated in Figure 5.2, based on LDA, where the input vocabulary V_i is redefined at each level i. The initial corpus C is input to the LDA model, consisting of M documents and a bag of unigrams taken from a vocabulary of unigrams $u_1 \in V_1$. LDA inference at level 1 results in a ranking of unigrams given topics at level 1, Φ_1, and topics given documents Θ_1. We concatenate the most probable unigrams given topics, Φ_1^T, with the set of location labels L to form bigrams as the new vocabulary to level 2, denoted $V_2 = \Phi_1^T L$. In cases where labels do not contain time information, the concatenation would be done with the initial vocabulary (i.e., $V_2 = \Phi_1^T V_1$) though here we avoid redundancy in time information in the vocabulary. The novelty of the method lies in the formulation of the n-gram vocabulary, which is guided by the topic outputs. More generally, the top ranked n-grams given topics, Φ_i^T at level i, are concatenated with the possible location labels to form the

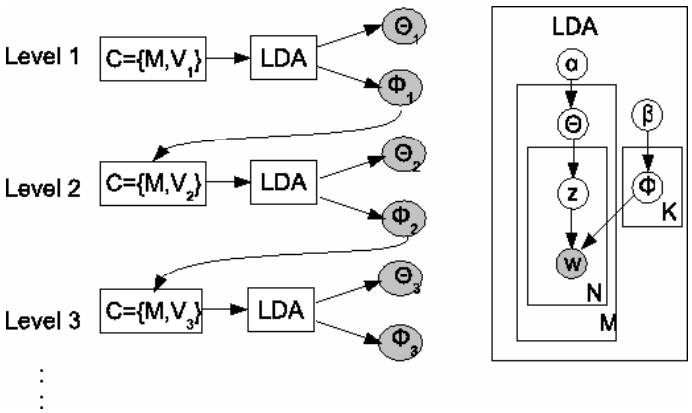


Figure 5.2. MLTM Overview. At level 1, the corpus C of M documents and words from the vocabulary of unigrams V_1 are input to the LDA model, whose graphical model is expanded on the right. The output from LDA at level 1 results in a ranking of unigrams given topics Φ_1 and topics given documents Θ_1. We concatenate the top ranked unigrams Φ_1^T with the possible location labels L to form bigrams as the new vocabulary to level 2, $V_2 = \Phi_1^T L$. This process can continue for n levels, resulting in output sequences of length n from the n-th LDA model.

new vocabulary for the next level of LDA $V_{i+1} = \Phi_i^T L$. Essentially, we are pruning and extending our vocabulary based on topic relevance and not only on simple frequency of occurrence of individual labels.

5.2.2 Experiments and Results

Dataset

For experiments, we remove days which contain entirely N (no reception) labels since they do not provide any useful information. The resulting dataset consists of 10118 days and over 242800 hours of data. This amounts to just over 21% of the days containing at least a single location label. For the multi-level LDA model, we set the number of topics K_i to 50 for all levels i. The LDA hyperparameters are set to $\beta_i = 0.01$ and $\alpha_i = 50/K_i$ for all levels i. These hyperparameters are chosen as in earlier chapters based on standard values used for text analysis [38] though experiments with other values of the same order produce the same results. 14 levels are considered for experiments, resulting in routines discovered based on vocabularies ranging from half-hour intervals (at level 1) to seven-hour intervals (at level 14).

Multi-Level Topics

The 50 topics at each level revealed human activities in terms of their locations for varying durations. The results are evaluated in terms of the most probable n-grams given topics and by most probable days given topics. We select several topics at various levels, and plot the 10 most probable days. We also plot the most probable n-grams Φ^T for the topics in addition to listing the two most probable n-grams in a separate table. In general, we observe that for most topics, as the level increases, the routines discovered occur over longer durations (as expected) though this duration is modeled by the vocabulary but is not

explicitly obtained as output. Further, as the level increases, the routines become more refined and discriminant over the day. These findings are explained in more detail in the paragraphs that follow.

Figure 5.3 illustrates results seen at various levels. The plots with the y-axis labeled 'Day' show the 10 most probable days for the topic, where the x-axis is the time of day and the possible location labels consist of $\{'W','H','O','N'\}$ as seen by the colorbar, where red is work, yellow is home, cyan is out and blue is no reception. The second, fourth, and sixth rows of Figure 5.3 visualize the most probable n-grams Φ_i^T for level i with $\Phi_i^T > 0.01$, where the x-axis is the time of day, the y-axis is the location label, and the value corresponds to the probability. From these plots, the n-grams duration cannot be determined, so we also include the corresponding n-grams in Table 5.1. Table 5.1 lists the 2 most probable n-grams for the topics in Figure 5.3, as well as the probability of the n-grams.

The MLTM discovers sequences of being at home, work, out as well as varying transitions between the possible locations at various times. Figure 5.3 (a) and (b) are level 1 routines of "at home in the morning". At level 2 (Figure 5.3 (d) and (e)) we can see some 'home-out-home' and 'work-out-work' sequences of activity. At level 8 (Figure 5.3 (f) to (j)), we discover "being at work or home or out" occurring over several hour intervals. At level 14 (Figure 5.3 (k) to (o)), the sequences discovered in the set of days occur over a large portion of the day and overall there is less variation in activity between the sets of top days discovered. In general, the set of n-grams discovered for higher levels (e.g. levels 8 and 14) are very discriminant sequences occurring over parts of the day. More specifically, level 8 Figure 5.3 (f) and (g) have 'work' followed by 'out with the transition always occurring at the same time. Figure 5.3 (h) is 'home' in the morning followed by 'out' occurring precisely at 9:30 am. The results show that increasing the sequence length of the input vocabulary can result in routine disambiguation as well as more precisely "filtered" output.

Comparing the results in terms of the actual n-grams per topic (Table 5.1) gives some temporal information about the sequences discovered. The top two words for topic 1 at level 1 Table 5.1 (a) are "home 5:30 - 6 am" and "home 6 - 6:30 am". Though these are unigrams, actually they often co-occur and are sequential in time. This can also be seen by the top documents in Figure 5.3 (a). In fact the plot of Φ_1^T shows a set of roughly 18 unigrams in a sequence defining this activity. However, these unigrams all have differing probabilities given the topic, with the last and first few unigrams in the sequence having lower probabilities. This results in the "noise" seen at the beginning and end of the "at home" activities mined from the data. To discriminate more between these activities over longer portions of the day, larger n-grams are necessary. For example, at level 8 topic 24 the "at home in the morning" routine is more specific with an 'out' transition occurring at 9:30 am.

Note that since topics are discovered for the most co-occurring words in a set of days and we do not explicitly model time durations in the MLTM, the routines do not necessarily correspond to exactly $n/2$ hours of a routine at level n as seen by the plots of the 10 most probable days. This is because topics capture n-grams co-occurring with other n-grams often overlapping in time. However, each level contains a set of topics which are unique and can not be discovered at other levels, revealing varying dominant routines.

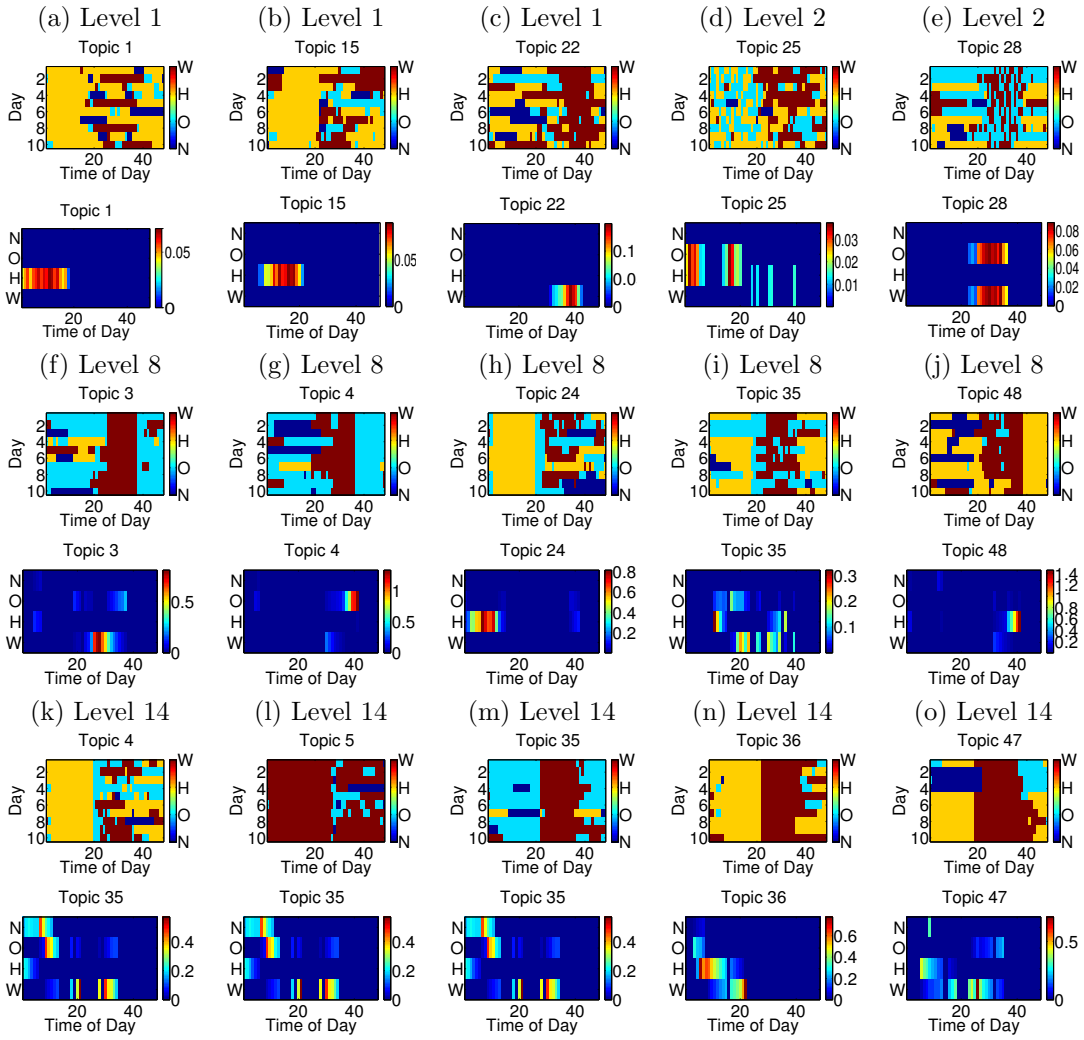


Figure 5.3. MLTM Results. Plots with the y-axis labeled 'Day' show the 10 most probable days for the topic, where the x-axis is the time of day. The second, fourth, and sixth rows visualize the most probable n-grams Φ_i^T, where the x-axis is the time of day, the y-axis is the location and the value corresponds to the probability of that location label occurring at that time. In general, the set of n-grams discovered for higher levels correspond to longer sequences of activities in terms of people's locations and are discriminant sequences occurring over longer intervals of the day.

Table 5.1. The two most probable location words which are the n-grams for a topic at level n. The topics labeled by letters correspond to those of Figure 5.3. For instance, in (a) the two most probable words for topic 1 at level 1 are "Home from 5:30 - 6 am" and "Home from 6 - 6:30 am". The 10 most probable days for this topic are plotted in Figure 5.3 (a).

(a) Level 1 - Topic 1

Home 5:30 - 6 am	0.071
Home 6 - 6:30 am	0.069

(b) Level 1 - Topic 15

Home 8 - 8:30 am	0.089
Home 6:30 - 7 am	0.087

(c) Level 1 - Topic 22

Work 7 - 7:30 pm	0.149
Work 7:30 - 8 pm	0.137

(d) Level 2 - Topic 25

Home 12 - 12:30 am followed by Out 12:30 - 1 am	0.024
Out 2 - 2:30 am followed by Home 2:30 - 3 am	0.021

(e) Level 2 - Topic 28

Out 3 - 3:30 pm followed by Work 3:30 - 4 pm	0.049
Work 5:50 pm followed by Out 5:30 - 6 pm	0.048

(f) Level 8 - Topic 3

Work 2 - 6 pm	0.101
Work 1 - 5 pm	0.096

(g) Level 8 - Topic 4

Out 8:30 - 12:30 pm	0.148
Out 8 - 12 pm	0.132

(h) Level 8 - Topic 24

Home 3:30 - 7:30 am	0.102
Home 5:30 - 9:30 am	0.094

(i) Level 8 - Topic 35

Home 5:30 - 9 am followed by Out 9 - 9:30 am	0.046
Home 6 - 9 am followed by Out 9 - 10 am	0.034

(j) Level 8 - Topic 48

Home 1:30 - 5:30 am	0.137
Home 12 - 4 am	0.136

(k) Level 14 - Topic 4

Home 12 - 7 am	0.145
Home 12:30 - 7:30 am	0.124

(l) Level 14 - Topic 5

Work 2 - 9 am	0.096
Work 2:30 - 9:30 am	0.095

(m) Level 14 - Topic 35

Out 10:30 - 11 am followed by Work 11 am - 5:30 pm	0.044
Work 3 - 9:30 pm followed by Out 9:30 - 10 pm	0.036

(n) Level 14 - Topic 36

Work 11 am - 6 pm	0.055
Home 3 - 10 am	0.053

(o) Level 14 - Topic 47

Work 1 - 7:30 pm followed by Out 7:30 - 8 pm	0.037
Home 3 - 9:30 am followed by Work 9:30 - 1 am	0.032

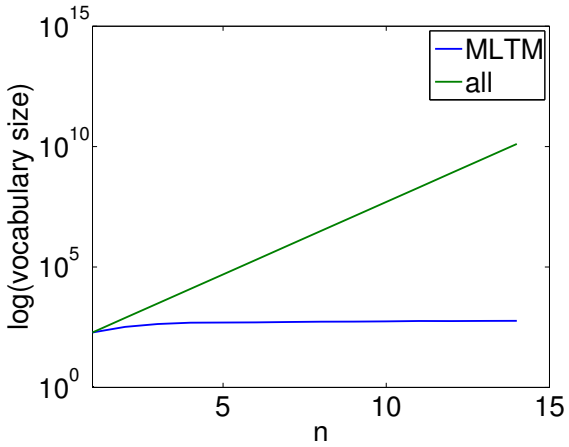


Figure 5.4. The growth in vocabulary size for the naive n-gram case ('all') versus our Multi-Level Topic Model (MLTM) for $\Phi_i^T > 0.01$. The naive case grows exponentially. In contrast, the MLTM vocabulary does not grow exponentially, and is very effective in limiting the n-gram vocabulary size at large n.

Vocabulary Size

We plot the growth in vocabulary over multiple levels for the case where $\Phi_i^T > 0.01$ in Figure 5.4. The logarithm of the vocabulary size is displayed as a function of the number of sequential labels n for the naive n-gram case (labeled 'all') and the Multi-Level Topic Model. In the naive n-gram case the growth in the vocabulary size is exponential, where the vocabulary size at level n is $xx^n * yy$, where $xx = 4$ and $yy = 48$ in this case. We take into account timeslots in constructing the time series sequences which results in a factor of 48. For 'MLTM' $V_i = \Phi_{i-1}^T L$. If we consider all the possible n-grams for which $n <=$ level, where at each level $V_i = (V_1, \Phi_1^T L, \Phi_2^T L, ...\Phi_{i-1}^T L)$, the vocabulary growth is approximately a cumulative sum of 'MLTM' in Figure 5.4 which is not exponential. This shows the advantage in constraining the vocabulary with our proposed approach. If LDA were to be used with an exponentially large vocabulary there would be problems due to sparsity, as well as computational constraints.

Vocabulary Analysis

We have established that the vocabulary size does not grow exponentially, now we analyze what we are actually capturing by comparing the vocabulary constructed by MLTM and the most frequently occurring n-grams. One question that arises is whether the T top n-grams Φ_i^T are not simply the most frequent n-grams. To investigate this we compute the percentage of overlapping n-grams between the T most frequently occurring n-grams and the n-grams formed as input to level n, $\Phi_{n-1}^T L$. In Figure 5.5, we plot the percentage of n-gram overlap for $T = 10$ and $T = 50$, considering n-grams for $n = 1$ to $n = 14$. The percentage of overlap for the 10 most occurring n-grams with the MLTM vocabulary is high, which shows we actually capture the most critical n-grams. For $n = 3$ to $n = 9$, we capture 9 out of 10 of the most frequently occurring n-grams, which is very good. When considering the 50 most frequently occurring n-grams in the data, we capture a large percentage of them for all n, which is critical

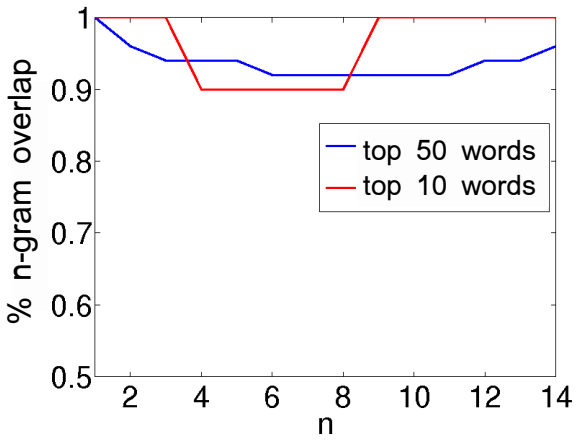


Figure 5.5. Percentage of overlap between the T most frequently occurring n-grams in the corpus and the vocabulary of MLTM n-grams at level n consisting of $\Phi_{n-1}^{T}L$. Results are shown for $T = 10$ and $T = 50$. The percentage of overlap for the 10 most occurring n-grams and the MLTM vocabulary is high, which shows we are capturing the most occurring n-grams with our method. Considering the 50 most frequently occurring n-grams in the data, the MLTM vocabulary contains a large percentage of them for all n, however, the MLTM vocabulary also captures a small percentage of other n-grams, which corresponds the difference between finding the most frequently occurring n-grams, and the most frequently co-occurring n-grams.

in showing that the method works effectively. However, we also capture some differing n-grams which is the difference between finding the dominating co-occurrences and not just the most frequently occurring labels.

Limitations

Though the multi-level topic model revealed meaningful results, it also has limitations. The first one is that at each level only a fixed number of consecutive words (or n-gram size) is considered. The experiments discussed here could easily be modified to have the i^{th} level vocabulary constructed as $V_i = (V_1, \Phi_1^T L, \Phi_2^T L, ...\Phi_{i-1}^T L)$, though this was not done since the vocabulary we chose, the results in terms of vocabulary size, sequences discovered and most frequently occurring n-grams captured revealed to be very good. Another limitation of this work is the pre-specification of the number of topics and hyperparameters at each level, though this is a limitation of topic models in general.

5.3 Pairwise-Distance Topic Model

In this section, we introduce a new generative model for long sequence generation. The model is built on Latent Dirichlet Allocation, with the extension of handling large sequences of n consecutive words (or location labels in this case) without an explosion of number of parameters nor parameter dimension. If we consider a sequence, $\mathbf{q} = (\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_n})$, to be a set of consecutive words, then the Pairwise-Distance Topic Model (PDTM) can generate a corpus of sequences. Sequences can be viewed as n-grams, and we use both terms throughout. The maximum length of the sequence n is predefined. This model was designed to manage large n without model parameter dimension explosion after $n = 3$, which is the

case in previous models [93]. As discussed earlier in the chapter, a word in an n-gram is assumed to be conditionally dependent on all previous words in the sequence, thus making large sequences infeasible to manage. In contrast, we integrate latent topics and assume a word in the sequence to be conditionally dependent on the first word as well as the topic and remove the dependence on all other words. We apply this model to location data to discover activities over large durations considering intervals of up to several consecutive hours. Next we define the generative process, derive the learning and inference procedure and present the results.

5.3.1 Data Representation

The PDTM can be applied to any type of data with discrete valued elements in a sequence, for example text, preprocessed video, or mobile sensor data. In this case we consider mobile location data over time. We make an analogy with LDA where a document is an interval of time in a person's daily life, for example a week or a day. A word $w = (t, l)$ is composed of a location $l \in L = \{'H', 'N', 'O', 'W'\}$ and a time coordinate of the day $t \in T = \{1, 2, 3, ..., tt\}$ where $tt = 48$ in this case.

5.3.2 The Probabilistic Model

The graphical model for our pairwise-distance topic model is illustrated in Figure 5.6. As before, we use a probabilistic approach where observations are represented by random variables, highlighted in gray in the figure. The latent variable z corresponds to a topic of activity sequences. The model parameters are defined in Table 5.2.

Table 5.2. Symbol description

n	The length of the sequence
$\mathbf{q}$	A sequence defined as n consecutive words (w_1, ..., w_n)
m	An instance of a document
S_m	The total number of sequences $\mathbf{q}$ in document m
M	The number of documents in the corpus
T	The number of latent topics
z	A latent topic
V	The vocabulary size
$\boldsymbol{\Theta}$	The distribution of topics given documents
$\boldsymbol{\Phi}$	The distribution of sequences given topics, where $\boldsymbol{\Phi} = \{\boldsymbol{\Phi}_{\mathbf{1_z}}, \boldsymbol{\Phi}_{\mathbf{2_{z,w_1}}}, ..., \boldsymbol{\Phi}_{\mathbf{n_{z,w_1}}}\}$
$\boldsymbol{\Phi}_{\mathbf{1_z}}$	The distribution of w_1 given topics
$\boldsymbol{\Phi}_{\mathbf{j_{z,w_1}}}$	The distribution of w_j given w_1 and topics for $1 < j \leqq n$

The generative process is defined as follows:

1. Initialization:

   (a) For each document m in the corpus draw a distribution over topics $\boldsymbol{\theta_m} \sim$ Dirichlet(α).

   (b) For each document m in the corpus:

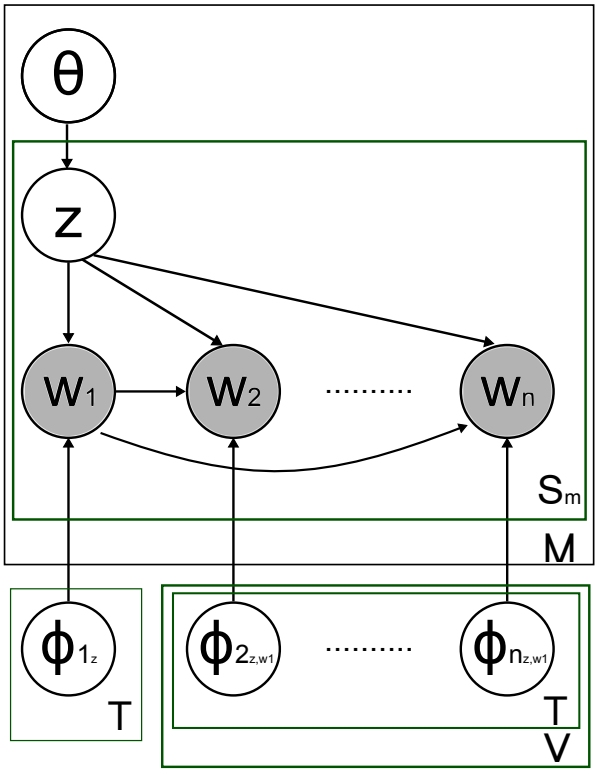


Figure 5.6. Graphical model of the Pairwise-Distance Topic Model (PDTM). A sequence $\mathbf{q}$ is defined to be n consecutive words $\mathbf{q} = (w_1, w_2, ..., w_n)$. This generative model is used for sequence generation based on an extension of LDA.

(c) For each sequence $\mathbf{q}$ of the S_m sequences in document m:

(d) Draw a distribution over words $\boldsymbol{\phi}_{\mathbf{1}_z} \sim \text{Dirichlet}(\beta_1)$ for each first word in the n-gram.

(e) For each consecutive word in the n-gram, w_j where $1 < j \leqq n$, draw a distribution over words $\boldsymbol{\phi}_{\boldsymbol{j}_{z,w_1}} \sim \text{Dirichlet}(\beta_j)$. Here $\boldsymbol{\phi}_{\boldsymbol{j}_{z,w_1}}$ captures the dependency with z as well as with w_1. Note n is defined by the user and is fixed.

2. Sequence generation procedure.

(a) For each document m in the corpus:

(b) For each sequence $\mathbf{q}$ of the S_m sequences in document m:

(c) Draw a topic $z|m \sim \text{Multinomial}(\boldsymbol{\theta}_{\boldsymbol{m}})$.

(d) Draw the first word in the sequence $w_1|z \sim \text{Multinomial}(\boldsymbol{\phi}_{\mathbf{1}_z})$.

(e) For $j = 2$ to n:

(f) Draw the j-th word in the sequence $w_j|w_1, z \sim \text{Multinomial}(\boldsymbol{\phi}_{\boldsymbol{j}_{z,w_1}})$ for $1 < j \leqq n$.

In summary, in the generative process for each sequence, the model first picks the topic z of the sequence and then generates all the words in the sequence. The first word in the sequence is generated according to a multinomial distribution $\boldsymbol{\phi}_{\mathbf{1}_z}$, specific to the topic z. The remaining words in the sequence, w_j for $1 < j \leqq n$, are generated according to a multinomial $\boldsymbol{\phi}_{\boldsymbol{j}_{z,w_1}}$ specific to the topic z as well as the first word of the sequence w_1. Note j is the j-th word in the sequence, but it can also be viewed as the distance between the word j and word 1.

We assume a Dirichlet prior distribution for $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi} = \{\boldsymbol{\Phi}_{\mathbf{1}_z}, \boldsymbol{\Phi}_{\mathbf{2}_{z,w_1}}, ..., \boldsymbol{\Phi}_{\mathbf{n}_{z,w_1}}\}$ with hyperparameters α and $\beta = \{\beta_1, \beta_2, ..., \beta_n\}$, respectively. We assume symmetric Dirichlet distributions with scalar parameters α and β such that $\alpha = \sum_{k=1}^{T} \frac{\alpha_k}{T}$, $\beta_1 = \sum_{v=1}^{V} \frac{\beta_{1,v}}{V}$, and $\beta_j = \sum_{t_1=1}^{V} \sum_{t_2=1}^{V} \frac{\beta_{(t_1,t_2)_j}}{V^2}$ for $1 < j \leqq n$. Note the parameters α_k, $\beta_{1,v}$, and $\beta_{(t1,t2)_j}$ are the components of the hyperparameters α, β_1, and β_j, respectively in the case of non-symmetric Dirichlet distributions. The joint probability of observations and latent topics can be obtained by marginalizing over the hidden parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$. These relations are then used for inference and parameter estimation:

$$
\begin{aligned}
p(\mathbf{z},\mathbf{q}|\alpha,\beta) &= p(\mathbf{z},\boldsymbol{w_1},...,\boldsymbol{w_n}|\alpha,\beta) && (5.1)\\
&= p(\boldsymbol{w_1},...,\boldsymbol{w_n}|\mathbf{z},\alpha,\beta)\cdot p(\mathbf{z}|\alpha,\beta) && (5.2)\\
&= p(\boldsymbol{w_2},...,\boldsymbol{w_n}|\boldsymbol{w_1},\mathbf{z},\alpha,\beta)\cdot p(\boldsymbol{w_1}|\mathbf{z},\alpha,\beta)\cdot p(\mathbf{z}|\alpha,\beta) && (5.3)\\
&= p(\mathbf{z}|\alpha)p(\boldsymbol{w_1}|\mathbf{z},\beta_1)\prod_{j=2}^{n} p(\boldsymbol{w_j}|\mathbf{z},\boldsymbol{w_1},\beta_j) && (5.4)\\
&= \int_{\boldsymbol{\Theta}} p(\mathbf{z}|\boldsymbol{\Theta})p(\boldsymbol{\Theta}|\alpha)d\boldsymbol{\Theta}\cdot\int_{\boldsymbol{\Phi_1}} p(\boldsymbol{w_1}|\mathbf{z},\boldsymbol{\Phi_1})p(\boldsymbol{\Phi_1}|\beta_1)d\boldsymbol{\Phi_1}\cdot\\
&\quad \prod_{j=2}^{n}\int_{\boldsymbol{\Phi_j}} p(\boldsymbol{w_j}|\boldsymbol{w_1},\mathbf{z},\boldsymbol{\Phi_j})p(\boldsymbol{\Phi_j}|\beta_j)d\boldsymbol{\Phi_j} && (5.5)\\
&= \prod_{m=1}^{M}(\frac{1}{B(\alpha)}\int\prod_{k=1}^{T}\boldsymbol{\theta}_{m,k}^{n_m^k+\alpha-1}d\boldsymbol{\theta})\cdot\prod_{k=1}^{T}(\frac{1}{B(\beta_1)}\int\prod_{t=1}^{V}\boldsymbol{\phi}_{1_{k,t}}^{n_k^t+\beta_1-1}d\boldsymbol{\phi}_1)\cdot\\
&\quad \prod_{j=2}^{n}\prod_{k=1}^{T}\frac{1}{B(\beta_j)}(\int\prod_{t_1=1}^{V}\prod_{t_2=1}^{V}\boldsymbol{\phi}_{j_{k,t_1,t_2}}^{n_{k_i^{`}}^{(t_1,t_2)_j}+\beta_j-1}d\phi_j) && (5.6)\\
&= \prod_{m=1}^{M}\frac{B(n_m+\alpha)}{B(\alpha)}\cdot\prod_{k=1}^{T}(\frac{B(n_k+\beta_1)}{B(\beta_1)}\cdot\prod_{j=2}^{n}\frac{B(n_{k_j^{`}}+\beta_j)}{B(\beta_j)}) && (5.7)
\end{aligned}
$$

where $p(\mathbf{z}|\alpha)$, $p(\boldsymbol{w_1}|\mathbf{z},\beta_1)$, and $p(\boldsymbol{w_j}|\boldsymbol{w_1},\beta_j)$ are derived in Appendix B resulting in the following.

$$
\begin{aligned}
p(\mathbf{z}|\alpha) &= \prod_{m=1}^{M}\frac{B(n_m+\alpha)}{B(\alpha)} && \text{where} \quad n_m=\{n_m^k\}_{k=1}^{T} && (5.8)\\
p(\boldsymbol{w_1}|\mathbf{z},\beta_1) &= \prod_{k=1}^{T}\frac{B(n_k+\beta_1)}{B(\beta_1)} && \text{where} \quad n_k=\{n_k^t\}_{t=1}^{V} && (5.9)\\
\text{and for} & \quad 1<j<=n\\
p(\boldsymbol{w_j}|\boldsymbol{w_1},\mathbf{z},\beta_j) &= \prod_{k=1}^{T}\frac{B(n_{k_j^{`}}+\beta_j)}{B(\beta_j)} && \text{where} \quad n_{k_j^{`}}=\{n_{k_j^{`}}^{(t_1,t_2)_j}\}_{t_1=1,t_2=1}^{V,V} && (5.10)
\end{aligned}
$$

We define the following notation; n_m^k is the number of occurrences of topic k in document m; $n_m=\{n_m^k\}_{k=1}^{T}$; n_k^t is the number of occurrences of term t in topic k, $n_k=\{n_k^t\}_{t=1}^{V}$; finally $n_{k_j^{`}}^{(t_1,t_2)_j}$ is the number of occurrences of terms $t_2$ occurring $j$ words after term $t_1$ in topic $k$ and $n_{k_j^{`}}=\{n_{k_j^{`}}^{(t_1,t_2)_j}\}_{t_1=1,t_2=1}^{V,V}$.

5.3.3 Inference and Parameter Estimation

Like LDA, the optimal estimation of model parameters is intractable. However, there are several techniques for inference including variational methods [10], Laplace approximation, and Markov Chain Monte Carlo methods (MCMC) [38]. Here we derive the model parameters based on the MCMC approach of Gibbs sampling [38]. We sample from the posterior distribution $p(\mathbf{z}|\mathbf{q})$ given the training data $\mathbf{q}$. In order to sample from $p(\mathbf{z}|\mathbf{q})$ using the Gibbs sampling method, we need to obtain the full conditional posterior distribution $p(z_i|\mathbf{z}_{-i}, \mathbf{q})$ where $\mathbf{z}_{-i}$ denotes all z_j's with $j \neq i$. The main steps in the formulation follow Equations 5.11-5.13, and a fully derived version is in Appendix B.

$$
\begin{aligned}
p(z_i = k|\mathbf{z}_{-i}, \mathbf{q}, \alpha, \beta) &= \frac{p(\mathbf{z}, \mathbf{q}|\alpha, \beta)}{p(\mathbf{z}_{-i}, \mathbf{q}|\alpha, \beta)} \qquad (5.11)\\
&= \frac{p(\mathbf{z}|\alpha)}{p(\mathbf{z}_{-i}|\alpha)} \cdot \frac{p(\boldsymbol{w_1}|\mathbf{z}, \beta_1)}{p(\boldsymbol{w_1}_{-i}|\mathbf{z}_{-i}, \beta_1) \cdot p(\boldsymbol{w_1}_i)} \cdot \prod_{j=2}^{n} \frac{p(\boldsymbol{w_j}|\boldsymbol{w_1}, \mathbf{z}, \beta_j)}{p(\boldsymbol{w_j}_{-i}|\mathbf{z}_{-i}, \boldsymbol{w_1}_{-i}, \beta_j) \cdot p(\boldsymbol{w_j}_i)}
\end{aligned}
$$

using Equations 5.8-5.10, and the knowledge that

$\mathbf{z}_{-i}$, or $\mathbf{w}_{x_{-i}}$ indicate that token i is excluded from the topic or word $\mathbf{w}_x$

$$
\propto \frac{B(n_m + \alpha)}{B(n_{m_{-i}} + \alpha)} \cdot \frac{B(n_k + \beta_1)}{B(n_{k_{-i}} + \beta_1)} \cdot \prod_{j=2}^{n} \frac{B(n_{k'_j} + \beta_j)}{B(n_{k'_{j,-i}} + \beta_j)} \qquad (5.12)
$$

Note the proportionality stems from the terms $\boldsymbol{w}_{1_i}$ and $\boldsymbol{w}_{j_i}$

$$
\propto (n^k_{m,-i} + \alpha) \cdot \frac{n^t_{k,-i} + \beta_1}{\sum_{t=1}^{V} n^t_{k,-i} + \beta_1} \cdot \prod_{j=2}^{n} \frac{n^{(t_1,t_2)_j}_{k,-i} + \beta_j}{\sum_{t_1=1}^{V} \sum_{t_2=1}^{V} n^{(t_1,t_2)_j}_{k,-i} + \beta_j} \qquad (5.13)
$$

where $n^{(y)}_x = n^{(y)}_{x,-i} + 1$ if $x = x_i$ and $y = y_i$

and $n^{(y)}_x = n^{(y)}_{x,-i}$ in other cases.

where $n_k = \{n^t_k\}_{t=1}^{V}$ and $n_{k'_j} = \{n^{(t_1,t_2)_j}_{k'}\}_{t_1=1,t_2=1}^{t_1=V,t_2=V}$. We use the properties $B(x) = \frac{\prod_{k=1}^{dim x} \Gamma(x_k)}{\Gamma(\sum_{k=1}^{dim x} x_k)}$, and $\Gamma(y) = (y-1)!$.

The model parameters can then be estimated as follows:

$$\boldsymbol{\theta}_m^k = \frac{n_m^k + \alpha}{\sum_{k=1}^{T}(n_m^k + \alpha)} \tag{5.14}$$

$$\boldsymbol{\phi}_{1,k}^t = \frac{n_k^t + \beta_1}{\sum_{t=1}^{V}(n_k^t + \beta_1)} \tag{5.15}$$

$$\boldsymbol{\phi}_{j,k}^{(t_1,t_2)_j} = \frac{n_k^{(t_1,t_2)_j} + \beta_j}{\sum_{t_1=1}^{V}\sum_{t_2=1}^{V}(n_k^{(t_1,t_2)_j} + \beta_j)} \tag{5.16}$$

// GOAL: Given a training corpus, α, β, T and n, estimate the parameters n_m^k, n_k^t, and $n_k^{(t_1,t_2)_j}$ for $j = 2$ to n from which we can determine the model parameters $\boldsymbol{\theta}_m^k$, $\boldsymbol{\phi}_{1,k}^t$, and $\boldsymbol{\phi}_{j,k}^{(t_1,t_2)_j}$.

// Initialization
1) Initialize the count parameters, $n_m^k = 0$, $n_k^t = 0$, $n_k^{(t_1,t_2)_j}$=0 for $j = 2$ to n.
2) **Iterate** over each sequence $\mathbf{q}$ in the corpus:
  3) Sample a topic k from $k \sim Mult(\frac{1}{T})$.
    4) Update the count parameters n_m^k, n_k^t, $n_k^{(t_1,t_2)_j}$ as follows $n_m^k = n_m^k + 1$, $n_k^t = n_k^t + 1$, $n_k^{(t_1,t_2)_j} = n_k^{(t_1,t_2)_j} + 1$ for $j = 2$ to n.

// Run the chain
5) **Iterate** over a large number of iterations (e.g. 1000):
  6) **Iterate** over each sequence:
    7) Decrement the current sequence and sequence elements' topic assignments as follows $n_m^k = n_m^k - 1$, $n_k^t = n_k^t - 1$, $n_k^{(t_1,t_2)_j} = n_k^{(t_1,t_2)_j} - 1$ for $j = 2$ to n.
    8) Sample a topic k for the sequence from $p(z = k|\mathbf{z}_{\neg i}, \mathbf{w}) \propto (n_{m,-i}^k + \alpha) \cdot \frac{n_{k,-i}^t + \beta_1}{\sum_{t=1}^{V} n_{k,-i}^t + \beta_1} \cdot \prod_{j=2}^{n} \frac{n_{k,-i}^{(t_1,t_2)_j} + \beta_j}{\sum_{t_1=1}^{V}\sum_{t_2=1}^{V} n_{k,-i}^{(t_1,t_2)_j} + \beta_j}$.
    9) Increment the new topic assignments as follows $n_m^k = n_m^k + 1$, $n_k^t = n_k^t + 1$, $n_k^{(t_1,t_2)_j} = n_k^{(t_1,t_2)_j} + 1$ for $j = 2$ to n.

// Compute model parameters
10) Estimate the unknown parameters as follows
$\boldsymbol{\theta}_m^k = \frac{n_m^k + \alpha}{\sum_{k=1}^{T}(n_m^k + \alpha)}$, $\boldsymbol{\phi}_{1,k}^t = \frac{n_k^t + \beta_1}{\sum_{t=1}^{V}(n_k^t + \beta_1)}$, and
$\boldsymbol{\phi}_{j,k}^{(t_1,t_2)_j} = \frac{n_k^{(t_1,t_2)_j} + \beta_j}{\sum_{t_1=1}^{V}\sum_{t_2=1}^{V}(n_k^{(t_1,t_2)_j} + \beta_j)}$, for $j = 2$ to n.

Figure 5.7. Gibbs Sampling Algorithm for the Pairwise-Distance Topic Model.

5.3.4 Experiments and Results

Dataset

For experiments in the Reality Mining dataset, we again removed days which contain entirely no reception (N) labels. We experimented with many values of T and plot selected results for $T = 20$. A range of values of T give similar results, the difference being that when T is small, the overall most occurring topics are discovered and when T is larger, more activities are found. We also vary hyperparameters from $\beta_j = 1$ to 0.01 and $\alpha = 1$ to 0.01. As long as β_j and α are smaller than the order of word/topic and document/topic counts, the results are more or less the same. We plot results for $\beta_j = 0.1$ for $1 \leqq j \leqq n$ and $\alpha = 0.1$. We consider up to $n = 14$ due to Matlab numerical constraints. This could likely be easily extended for larger n with relatively simple numerical optimization techniques.

Activity Sequences

We visualize a set of 15 topics corresponding to activity sequences for various n. Figure 5.8 (a) to (o) correspond to dominant sequences discovered for 4 different values of n, $n = 3$ (Figure 5.8 (a)-(e)), $n = 5$ (Figure 5.8 (f)-(h)), $n = 9$ (Figure 5.8 (i)-(j)), and $n = 13$ (Figure 5.8 (k)-(o)). In Figure 5.8, we plot the results in terms of the 20 most probable days given topics, $p(z|d)$. In general, we can see emerging location patterns discovered within subsets of days in the corpus. For example, in Figure 5.8 (a) there is no reception in the morning. In (b) there is work after roughly 10 am, with out several hours later, followed by work again. As n increases, we generally discover longer duration location patterns, which are now defined in the output parameters of the model as will be explained in more detail in the tables and figures that follow. We find in general the topics reveal activities of a similar sort as Chapter 3 based on the top documents for topics. This visualization of the most probable documents given topics discovered reveals that our model does mine the data in a manner similar to LDA, which is reassuring. For example, we can see a home-work-home routine in Figure 5.8 (d). However, our proposed model has a great advantage of outputting precise details of the labels composing the sequences.

In Figure 5.9 we display the topic results using the pairwise-distance topic model in terms of the most probable sequence components given topics. We show both visually and in terms of the actual output, the top ranked sequence components given topics. For $n = 3$, the sequence is as follows $q = (w_1, w_2, w_3)$ and the results $w_2|w_1, z$, $w_3|w_1, z$ as well as their probabilities are given in tables in Figure 5.9. The figures in the first column of Figure 5.9 simply serve as a visualization of the overall sequences discovered. The first block of the figure is the first component of the sequence given the topic, $w_1|z$, where the y-axis corresponds to the possible words. Note, there are too many word possibilities to show clearly on the axis. The plot simply serves as an approximate time of day display. The actual sequence components can be seen in the accompanying tables next to the figures. The figures in the second column are illustrations of $w_2|w_1, z$ and $w_3|w_1, z$, where the y-axis is again w_1 (same as the first block), and the x-axis is also possible words, but without a time component. Note each sequence has only one time component, which we plot with w_1 in the x-axis. So w_2 and w_3 are on the x-axis. The tables show the actual pairwise-distance topic

model results with the probabilities. You can see the sequence O-O-O starting at 8 pm is discovered in (a) for Topic 3 ($n = 3$). Note O-*-O indicates that w_1 is O, followed by O in position 3 of the sequence (w_3), with any possible location label in position 2 (w_2). For the results with $n > 3$, shown in Tables 5.3, and 5.4, we simplify the results presented by just listing the sequences. So for $n = 3$, Topic 3, we would show this as 8 pm O-O-O and 5 am N-N-N.

Tables 5.3 and 5.4 show the sequences that defined the topics displayed in Figure 5.8. We display the two most probable sequences $q = (w_1, w_2, ..., w_n)$ for the topic, and in many cases these are split in time. For example, in Table 5.3 (f) the outputs $w_i|w_1, z$ and $w_i|w_{1'}, z$ do not necessarily have the same w_1, or $w_1 \neq w_{1'}$. In this case, the sequences are all displayed on separate lines, but they have the same sequence number. This is one limitation of the model, which is expanded on in the Limitations section. It can also be viewed as an advantage of the model, since we can capture varying duration sequences. For large n, even if a sequence of exactly length n is not present in the data, the model still returns topics with co-occurrence occurring with length n, though these do not necessarily have to be in sequence. For example, for $n = 13$ and Topic 10 seen in (n), the most probable sequence consists of W-W-W-W-W-W starting at 4 pm and O-(5*)-O-O-O-O-O-O-O starting at 4 am. Since there was no dominant 13-gram, a 6-gram is discovered to co-occur with another large sequence. Three hours of being at work starting from 4 pm often co-occurs with being out for 30 minutes at 4:30 am, then being out 2.5 hours later (7 am) for 3.5 consecutive hours. We observe the activities in general are similar to those in Chapter 3, but now we obtain the sequence information as output.

The relaxation of the sequence component dependency can be viewed as an advantage of the model which is able to handle noisy data. By only consider pairwise components in the sequence, and removing the dependency on the other components, we can tolerate a certain amount of noise occurring within the sequence. The model inherently allows for sequence discovery with noise, which is important for this type of data collection.

In Figure 5.10, we plot the perplexity of the PDTM over varying number of topics computed on 20% unseen test data. The experiments are conducted for a sequence length of $n = 8$. We can see the perplexity drops to a minimum at around 50 topics. We therefore use 50 topics in order to compare the PDTM performance to LDA, plot in Figure 5.11. Also, the perplexity results illustrate that for a large number of topics, the model does not overfit the data, since the perplexity does not increase, but remains stable.

In order to compare our PDTM to LDA, we adapt the vocabulary used for LDA to have a comparable format to that used in the PDTM. The vocabulary we use for LDA consists of a pair of locations, a timeslot, as well as the distance between the locations. This results in a very similar vocabulary structure, so that we can compare the model performance itself. The log-likelihood results on 20% unseen test data, are plotted in Figure 5.11. We plot the log-likelihood, averaged over all the test documents. We do not compare the results in terms of perplexity since the normalizing factors would be different, producing a bias towards LDA. The log-likelihood results reveal that for small n, LDA performs slightly better. However, as n increases, the PDTM consistantly has better generalization performance.

Table 5.3. Most probable topics discovered using the pairwise-distance topic model, presented in terms of the sequences output by the model. Results are shown for various sequence lengths n corresponding to those illustrated in Figure 5.8.

(f) $n = 5$ Topic 2

Sequence 1	6 am	H-H-H-H-H
Sequence 2	9 am	H-H
Sequence 2	10 am	O-*-O-O-O

(g) $n = 5$ Topic 14

Sequence 1	7 am	N-N
Sequence 1	3 am	O-*-O-*-O
Sequence 1	1:30 am	O-*-*-O
Sequence 2	1 am	O-O-*-O-O
Sequence 2	7 am	N-*-N

(h) $n = 5$ Topic 17

Sequence 1	10 pm	O-O-O
Sequence 1	4:30 am	H-*-*-*-H
Sequence 2	8:30 pm	O-O
Sequence 2	4:30 am	H-*-H-H
Sequence 2	10 pm	O-*-*-*-O

(i) $n = 9$ Topic 5

Sequence 1	3:30 pm	W-W-W-W-W-W-W-W
Sequence 2	5:30 am	H-H-H-H-H-H-H-H
Sequence 2	5:30 pm	O-*-*-*-*-*-*-*-O

(j) $n = 9$ Topic 17

Sequence 1	4:30 am	H-H-H-H-H-H-H-H-H
Sequence 2	6:30 am	H-H-H-H-H
Sequence 2	3 am	H-*-*-*-*-*-H-H-H

Table 5.4. Continuation of Table 5.3. The results in this table are for $n = 13$.

(k) $n = 13$ Topic 2

Sequence 1	9 am	H-H-H-H-H-H-H-W-W-W
Sequence 2	5 pm	N-N-N-N-N
Sequence 2	9 am	H-*-*-*-*-W-W-H-H-H

(l) $n = 13$ Topic 3

Sequence 1	3 pm	W-W-W-W-W-W-W
Sequence 1	1:30 pm	W-*-*-*-*-*-*-W-W-W-W-W
Sequence 1	4:30 am	O-*-*-*-*-*-*-*-*-*-*-O
Sequence 2	1:30 pm	W-W-W-W-W-W-*-*-*-*-*-W
Sequence 2	3 pm	W-*-*-*-*-*-W-W-W-W
Sequence 2	4:30 am	O-*-*-*-*-*-*-*-*-*-O-O

(m) $n = 13$ Topic 8

Sequence 1	4:30 am	H-H-H-H-H-H-H-H-H-H-H
Sequence 1	10:30 am	W-*-*-*-*-*-*-*-*-*-W
Sequence 1	12:30 am	O-*-*-*-*-*-*-*-*-*-*-O
Sequence 2	12:30 am	H-H-H-H-H-*-*-*-*-*-*-H
Sequence 2	7:30 am	O-*-*-*-*-*-*-O
Sequence 2	10:30 am	O-*-*-*-*-*-*-*-O
Sequence 2	10:30 am	W-*-*-*-*-*-*-*-*-W-W-*-W

(n) $n = 13$ Topic 10

Sequence 1	4 pm	W-W-W-W-W-W
Sequence 1	4 am	O-*-*-*-*-*-O-O-O-O-O-O-O
Sequence 2	4 am	O-O-O-O-O-O
Sequence 2	4 pm	W-*-*-*-*-*-W
Sequence 2	5 am	O-*-*-*-*-*-*-O-O-O-O-O-O

(o) $n = 13$ Topic 12

Sequence 1	2 pm	W-W-W-W-W-W-W-W-W-W
Sequence 1	noon	W-*-*-*-*-*-*-*-*-W-W-W
Sequence 1	1:30 am	H-*-*-*-*-*-*-*-*-*-*-H
Sequence 2	noon	W-W-W-W-W-W-W-W-W-*-*-*-W
Sequence 2	2 pm	W-*-*-*-*-*-*-*-*-W-W
Sequence 2	1:30 am	H-*-*-*-*-*-*-*-*-*-*-H

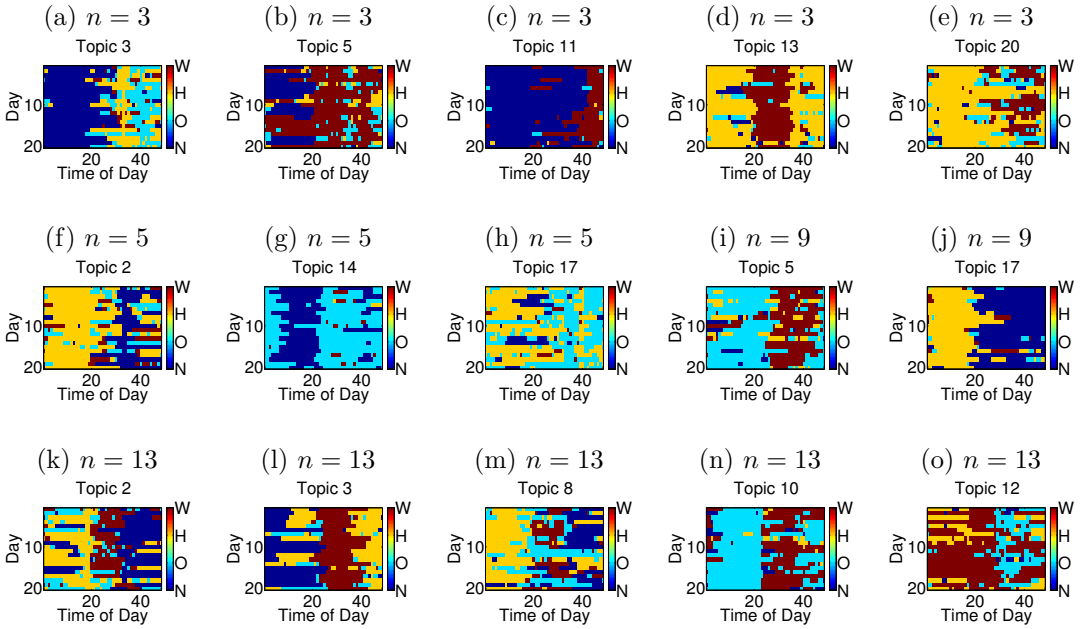


Figure 5.8. Topics discovered using the PDTM with various sequence lengths n. We plot the results in terms of the 20 most probable days given topics, $p(z|d)$. In general, we can see emerging location patterns discovered within subsets of days in the corpus. As n increases, we generally discover longer duration location patterns, which are now defined in the output parameters of the model as will be explained in more detail in the tables and figures that follow.

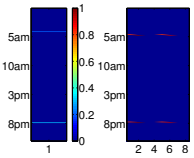
5.3.5 Limitations

There are two main limitations of the pairwise-distance topic model. The first limitation is that there is no constraint forcing the output components to be in sequence. More specifically, a valid output could be $w_2|w_1, z$ and $w_3|w_{1'}, z$ where $w_1 \neq w_{1'}$. This would not result in a sequence of length 3 since the first words differ. This is a limitation of the model in that the output can be out of sequence. However, it can also be an advantage in that the output produces varying length sequences. To solve this problem, we would have to add some constraints to the model. Another limitation is that output can contain overlapping components. For example, the top sequence output for a topic may be something like 3:30 pm H-H and 3 pm H-*-H. In that example, the components are overlapping in time. To address this problem, again, some constraints should be imposed regarding the time component in the word construction.

5.4 Comparison

We present an overall comparison of the two models developed. We compare them conceptually, and in terms of limitations, advantages, and disadvantages in Table 5.5.

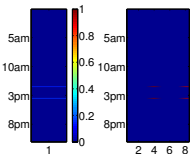
(a) Topic 3



(a) $n = 3$ Topic 3

$w_2\|w_1$	$p(w_2\|w_1)$	$w_3\|w_1$	$p(w_3\|w_1)$
8 pm O-O	0.25	8 pm O-*-O	0.23
5 am N-N	0.21	5 am N-*-N	0.21

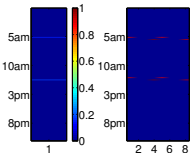
(b) Topic 5



(b) $n = 3$ Topic 5

$w_2\|w_1$	$p(w_2\|w_1)$	$w_3\|w_1$	$p(w_3\|w_1)$
3:30 pm W-W	0.15	3:30 pm W-*-W	0.14
1:30 pm W-W	0.13	1:30 pm W-*-W	0.12

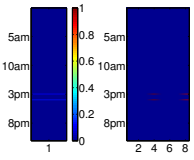
(c) Topic 11



(c) $n = 3$ Topic 11

$w_2\|w_1$	$p(w_2\|w_1)$	$w_3\|w_1$	$p(w_3\|w_1)$
12:30 pm W-W	0.16	12:30 pm W-*-W	0.15
5:30 am N-N	0.14	5:30 am N-*-N	0.14

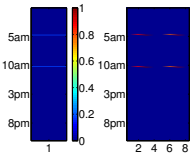
(d) Topic 13



(d) $n = 3$ Topic 13

$w_2\|w_1$	$p(w_2\|w_1)$	$w_3\|w_1$	$p(w_3\|w_1)$
4 pm W-W	0.12	4 pm W-*-W	0.11
3 pm W-W	0.11	3 pm W-*-W	0.11

(e) Topic 20



(e) $n = 3$ Topic 20

$w_2\|w_1$	$p(w_2\|w_1)$	$w_3\|w_1$	$p(w_3\|w_1)$
5 am H-H	0.17	5 am H-*-H	0.17
10:30 am O-O	0.15	10:30 am O-*-O	0.13

Figure 5.9. Topics discovered using the PDTM with various sequence lengths n, expressed in terms of the most probable sequence components for topics. We show both visually and in terms of the actual output, the top ranked sequence components given topics. The figures in the first column serves as a visualization of the overall sequences discovered. The first block of the figure is $w_1|z$, where the y-axis corresponds to the possible words. Since there are too many words to illustrate, the plot simply serves as an approximate time of day display. The actual sequence components can be seen in the accompanying tables next to the figures. The figures in the second column are illustrations of $w_2|w_1, z$ and $w_3|w_1, z$, where the y-axis is the same as the first block, and the x-axis is also possible words, but without a time component. The tables show the actual pairwise-distance topic model results with the probabilities.

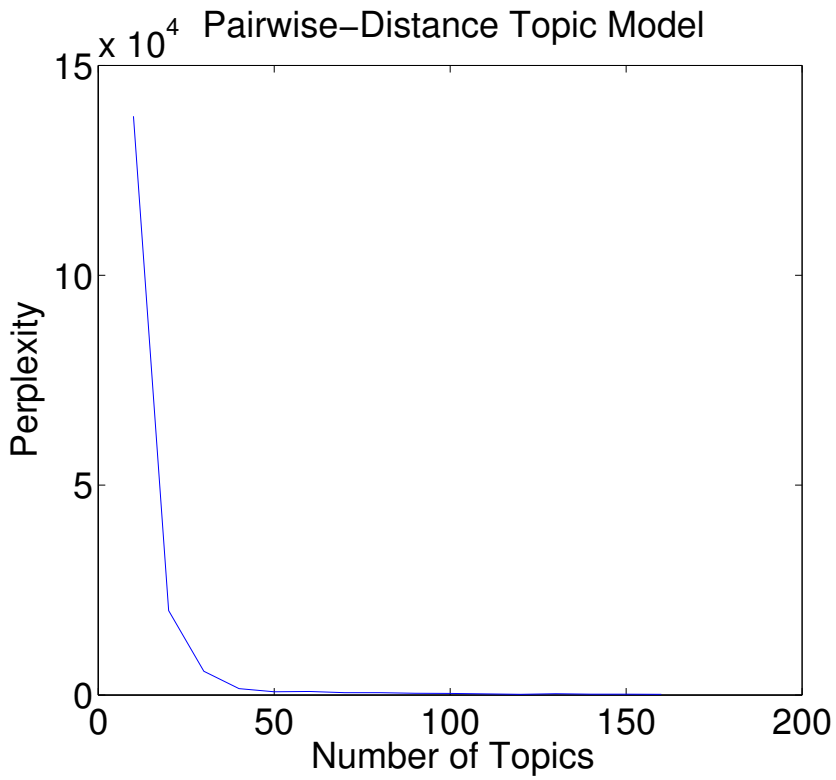


Figure 5.10. Perplexity of the PDTM over the number of topics on 20% unseen days (documents).

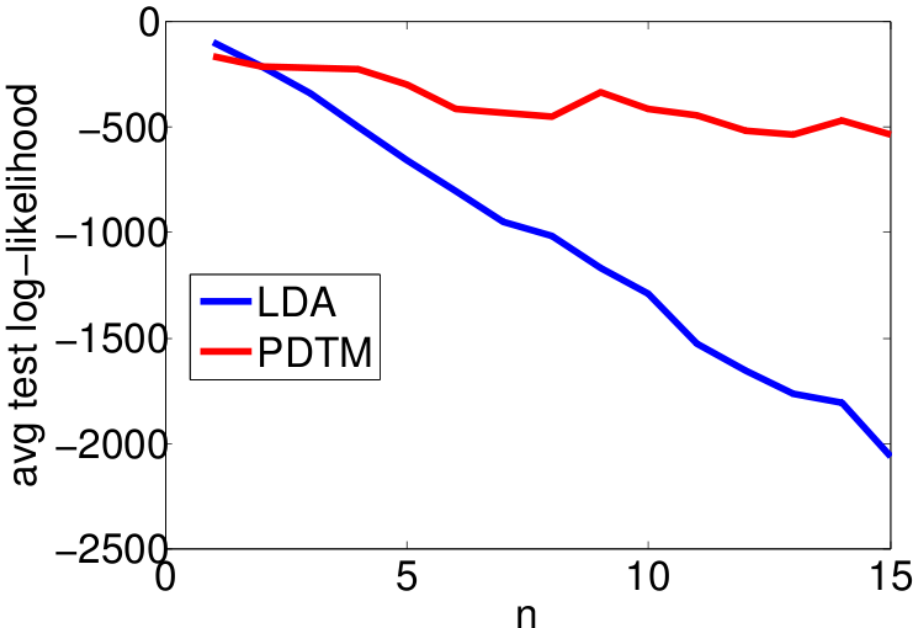


Figure 5.11. Average loglikelihood of the PDTM versus LDA on 20% unseen days (documents).

Table 5.5. Comparison of the pairwise-distance topic model and the multi-level topic model.

	pairwise-distance topic model	multi-level topic model
problem addressed	incorporating time in the model	incorporating time in the model
problem addressed	long duration sequence modeling	multiple duration sequence modeling
approach	pairwise modeling of words in sequences	multiple levels of LDA where the input vocabulary is changed based on the previous level's output
conceptual	non-iterative learning	iterative learning
predefined parameters	hyperparameters, number of topics, and max n	hyperparameters, number of topics at each level
advantages	handles noise in sequences	output with many simultaneous time durations
advantages	removes the need for a coarse-timeslot	removes the need for a coarse-timeslot
limitations	overlapping sequences in time and output sequences not constrained to have length n	number of topics at each level is not optimized
possible extensions	to impose constraints that handle overlapping sequences in the output and that restrict the output sequence to have length n	to introduce new techniques, perhaps hierarchical dirichlet processes, to learn the number of topics at each level automatically

5.5 Conclusion

In this chapter, we devise two models to address the limitations of activity modeling in terms of time constraints. The first model is called the multi-level topic model and it is designed to overcome the limitation of having coarse grain timeslots in the vocabulary. The model can output activities with multiple time durations, removing the need for previously defined timeslots in activity recognition. The second model is the pairwise-distance topic model and it is designed to enable the modeling of long sequences or long n-grams, which in turn also removes the need for predefined timeslots in the input features.

There are several future directions for both models proposed. For the multi-level topic model experiments could be performed with multiple n-gram sizes as vocabularies as the level increases. More specifically, at level 2, the input vocabulary would consist of unigrams and bigrams, at level 3 the vocabulary would consist of unigrams, bigrams, and trigrams, etc. Further investigations should be done on the disadvantages of iterative approaches. Perhaps there are non-iterative approaches to run inference on the model. Finally, as mentioned before, optimizing the number of topics at each level would greatly improve the model, especially if used with other applications.

For the pairwise-distance model, we plan to extend the work and consider dependency between a subset of previous words in the sequence, as opposed to just the first word in the sequence. This approach would

also be beneficial over the standard n-gram approach in considering dependency over all of the previous words, causing parameter size explosion very quickly. The problem then becomes how to determine the number of previous words for dependency, which we plan to approach with a hierarchical dirichlet process. Another future work is to apply both models to other types of data, perhaps containing ground truth for further objective evaluations.

Chapter 6

Modeling Opinion Change with Topic Models

An important question in the social sciences, as well as the practical arts of education, sales, and politics, is how ideas, opinions, innovations, and recommendations spread through society. Diffusion is the phenomena of propagation of ideas or opinions within a social network. Many important attributes of our lives are expressed primarily in real-world, face-to-face interactions. To model the adoption of these behaviors, we need fine-grained data about face-to-face interactions between people, i.e., who talks to whom, when, where, and how often, as well as data about exogenous variables that may affect the adoption process. Such social sensing of face-to-face interactions that explain social diffusion phenomena is a promising new area in pervasive computing.

Traditionally, social scientists have relied on self-report data to study social networks, but such approaches are not scalable. It is impossible to use these methods with fine resolution, over long timescales (e.g., months or years), or for a large number of people, (e.g., hundreds or thousands). Further, while people may be reasonably accurate in their reports of long term social interaction patterns, it is clear that memory regarding particular relational episodes is quite poor. As an example, in a survey of informant accuracy literature, Bernard et al. have shown that recall of social interactions in surveys used by social scientists is typically 30-50 % inaccurate [3, 12].

A key question is 'what influences opinion change'? Is there an underlying mechanism resulting in the change of opinion for some people? Can we measure this mechanism, and if so, can we predict future opinion changes from observed behavior? In this chapter as an extension to the work described in Chapters 3-5, we propose a method for modeling opinions of subjects obtained by survey questionnaire data and correlate this with user interaction and communication data obtained by mobile phone sensor data. The method used is based on the Latent Dirichlet Allocation (LDA) [10] topic model, to contrast the activities of participants that change opinions, with those who do not. The method discovers, in an unsupervised manner, the dominating routines of people in the dataset, where routines are the most

frequently co-occurring political opinion exposure patterns also referred to as topics in the rest of the chapter.

In this chapter we analyze data captured by Anmol Madan and Alex Pentland at the MIT Media Lab during the last three months of the 2008 US Presidential campaign. The dataset is from September 2008 to November 2008 during the presidential campaign of John McCain and Barack Obama and was collected amongst the residents of an undergraduate residence hall at a North American university. This dataset consists of 132,000 hours of social interactions data measured by mobile phones, and the political opinions measured using monthly surveys. Using an LDA-based topic modeling approach, we study the behavior differences between individuals who change opinions, and those who held their political opinions. We show statistically significant differences in the activities of people who changed their preferred party versus those who did not. People that changed preferred party often discuss face-to-face with their democrat political discussants, and their daily routines included heavy phone and SMS activity. We also find that people who decrease their interest in politics often interact with people that have little or no interest in politics.

This chapter is organized as follows. We first present the data characteristics including details about the questionnaires as well as some statistics of the sensor data. We then present the details of the topic modeling approach for opinion change modeling. The model selection is then presented followed by experimental results. The material presented here was done in collaboration with Anmol Madan and Alex Pentland and is published in [66].

6.1 Dataset Characteristics

In the past, researchers have used Call Data Records provided by mobile operators to better understand human behavior [37, 27]. Our approach, however, is to use pervasive sensing methods for capturing social interactions, and this has several advantages. Firstly, it allows us to sample different sensors and dependent training labels, and not just calling data alone. Secondly, from a privacy perspective, this requires the user's explicit participation in data collection. Additionally, in the future, it could be used to provide the user immediate feedback on the mobile device itself.

6.1.1 Sensors and Data

The dataset was recorded with Windows Mobile 6.x devices. The data contains the Bluetooth devices in proximity as well as the WLAN Access Point Identifiers (WLAN APs), which were scanned every 6 minutes. Call and SMS log details are recorded every 20 minutes, including information about missed calls and calls not completed. Periodic scanning of Bluetooth and WLAN APs reduced operational battery life with average usable life between 14-24 hours. More details can be found in [18].

The data was collected in a university undergraduate dormitory for an academic year, with a total of over seventy participants. This data was collected for several behavioral studies, including an epidemiology study [64] and an obesity study [65]. However, the data relevant for political opinions was collected

over a three-month period, from September to November 2008. There are 78 participants in the political opinion study, from which 36 are female and 42 are male. The participants birth year ranges from 1981 to 1991. The undergraduates are from all years, with 10 juniors, 22 freshman, 21 sophomore, 15 senior, and 10 others. The participants represent eighty percent of the total population of the dormitory. The remaining twenty percent of residents declined to participate in this study citing privacy concerns. The undergraduate dormitory is known for its pro-technology orientation and tight-knit community.

The mobile phone interaction dataset consists of approximately 450,000 bluetooth proximity scans, 1.2 million WLAN access-point scans, 16,900 phone call records and 17,800 SMS text message events. The average duration of phone calls is approximately 138 seconds, and 58 percent of phone calls were during weekdays.

6.1.2 Political Opinion Questionnaires

The subjects' political opinions were captured using three monthly web-based surveys, once each in September, October, and November 2008 (immediately following the presidential election). The monthly survey instrument was based on established political science literature, and consisted of questions shown in Table 6.1. The questions were identical to the survey instrument used by Lazer and Rubineau [56], who measured the monthly political opinions of students across different universities (during the same 2008 election period) and studied the co-evolution of political opinions and self-report friendship networks.

Political scientists have established that shifts in political opinions are gradual [46]. This is observed in our dataset, as approximately 30% of the participants changed their opinions for each of the dependent questions during the three-month observation period. Opinion changes were along 1-point or 2-points on the respective 4/7-point Likert scales. Similar variations in our dependent variables were also reported in the analysis of Lazer and Rubineau [56].

For each monthly survey, participants also identified other residents that were political discussants, close friends or social acquaintances, identical to those used in [56]. Baseline information including race, ethnicity, political opinions of the person's parents and religious affiliations was also collected before the start of the experiment.

We plot the survey responses for four of the survey questions, interest in politics (4-point scale), liberal or conservative (7-point scale), preferred party (4-point scale) and preferred party details (7-point scale) in Figure 6.1. The possible responses are as seen in the colorbars as follows

- {'NI', 'SI', 'I', 'VI', 'NR'} = {'not interested', 'slightly interested', 'somewhat interested', 'very interested', 'no response'}
- {'EC', 'C', 'SC', 'M', 'SL', 'L', 'EL', 'NR'} = {'extremely conservative', 'conservative', 'slightly conservative', 'moderate', 'slightly liberal', 'liberal', 'extremely liberal', 'no response'}
- {'Ot', 'Ind', 'Rep', 'Dem', 'NR'} = {'other party', 'independent', 'republican', 'democrat', 'no response'}
- {'SR', 'NSR', 'Rep', 'N', 'Dem', 'NSD', 'SD', 'NR'} = {'strong republican', 'not strong republican', 'republican', 'neither', 'democrat', 'not strong democrat', 'strong democrat', 'no response'}.

For the experiments that follow, we consider a change of opinion to occur if the subjects' opinion in November differs from their opinion in September. The statistics for number of users who changed opinions versus those who did not are shown in Table 6.2. There are 78 subjects in the study, however we only consider subjects that responded in both September and November. The number of participants for each opinion is shown in the Total Responses column of Table 6.2.

Table 6.1. Political Survey Instrument used to capture different political opinions. All responses were constructed as Likert scales.

Survey Question	**Possible Responses**
Are you liberal or conservative?	7-point Likert scale Extremely conservative to extremely liberal
How interested are you in politics?	4-point Likert scale Not interested to very interested
What is your political party preference?	7-point Likert scale Strong Democrat to strong Republican
Which candidate are you likely to vote for? (Sept and Oct)	Choice between leading Republican and Democrat nominees
Which candidate did you vote for? (Nov)	Choice between B. Obama and J. McCain
Are you going to vote in the upcoming election? (Sept and Oct)	4-point Likert scale
Did you vote in the election? (Nov)	Yes or No

Table 6.2. Statistics of numbers of users in various groups.

Opinion	Changed Opinion	Did not Change Opinion	Total Responses
interest in politics	14	34	48
liberal/conservative	17	32	49
preferred party	5	44	49
preferred party details	13	36	49

Basic Statistics of Mobile Sensor Data

In Figure 6.2 we plot some basic statistics on the call activities and interaction data for all users that responded to the interest in politics question in the survey. Plot (a) is a histogram of the call duration

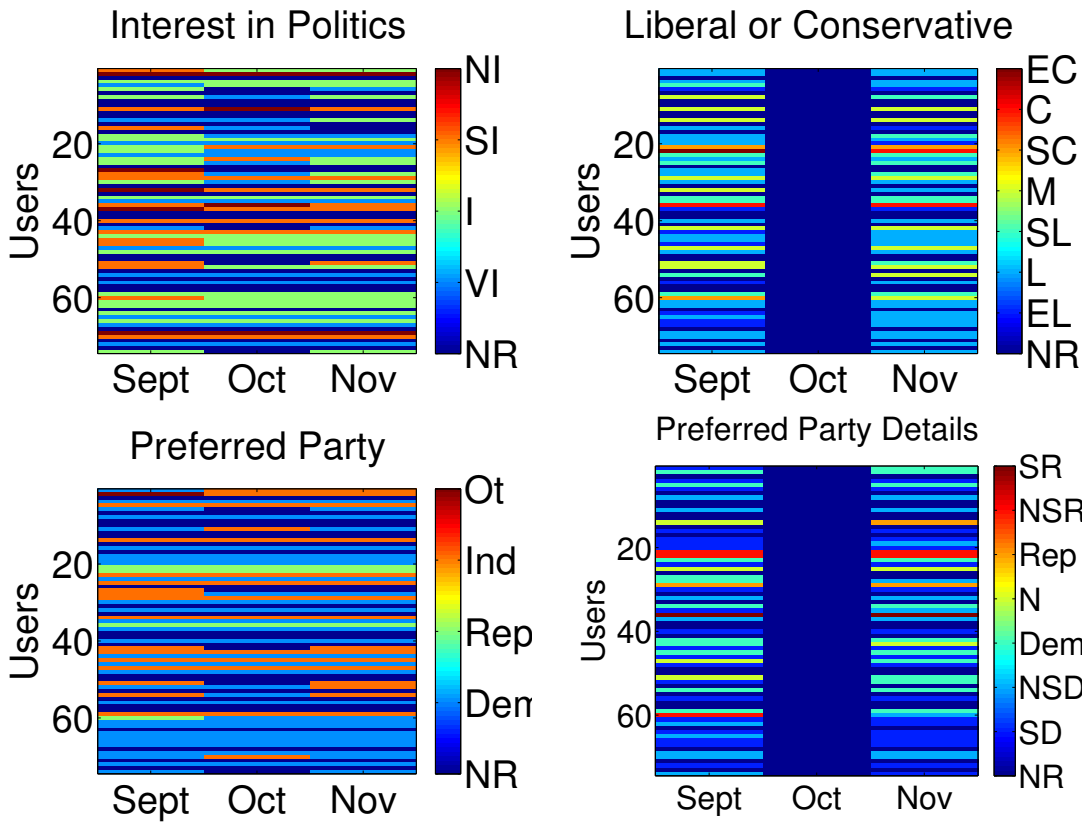


Figure 6.1. User opinions for 4 of the survey questions. The legends correspond to the following opinions, {'NI', 'SI', 'I', 'VI', 'NR'} = {'not interested', 'slightly interested', 'somewhat interested', 'very interested', 'no response'} {'EC', 'C', 'SC', 'M', 'SL', 'L', 'EL', 'NR'} = {'extremely conservative', 'conservative', 'slightly conservative', 'moderate', 'slightly liberal', 'liberal', 'extremely liberal'} {'Ot', 'Ind', 'Rep', 'Dem', 'NR'} = {'other party', 'independent', 'republican', 'democrat', 'no response'} {'SR', 'NSR', 'Rep', 'N', 'Dem', 'NSD', 'SD', 'NR'} = {'strong republican', 'not strong republican', 'republican', 'neither', 'democrat', 'not strong democrat', 'strong democrat', 'no response'}.

where we consider two groups, less than or equal to 10 minutes and greater than 10 minutes. We can see that most of the phone calls are less than or equal to 10 minutes in duration, especially those of individuals who are very interested in politics. In (b), we also consider histograms of number of SMS sent and received, grouped by interest in politics. Overall the group 'somewhat interested' sends the most SMS and 'not at all interested' sends the fewest. In Figure 6.2 (c) we consider histograms of Bluetooth (face-to-face) interactions grouped by interest. For interactions of less than or equal to 10 min, $11 - 45$ min, and $46 - 120$ min the counts are more of less evenly distributed. For interactions occurring over 2 hours there are more occurrences for subjects that have higher interest in politics. Next we present the construction of the vocabulary used for modeling opinion change. Note the features include the bluetooth and communication features plotted here in addition to survey responses on relationships and opinions.

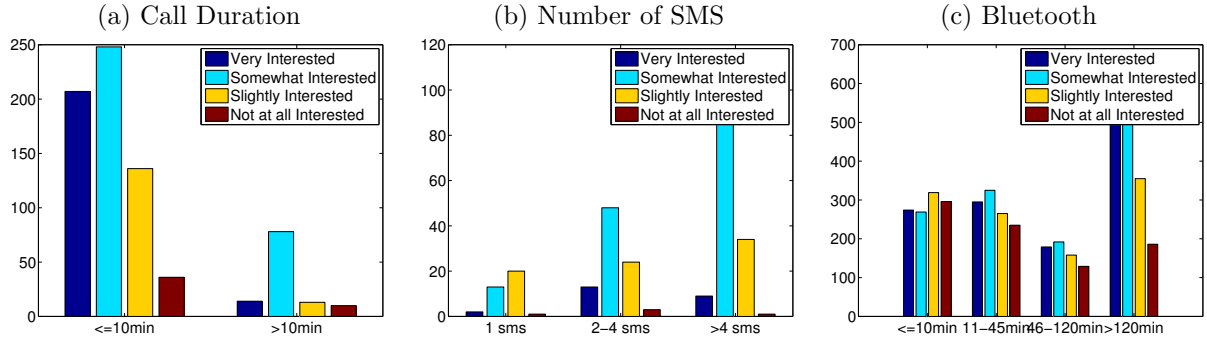


Figure 6.2. User communication statistics grouped according to the interest in politics opinion.

6.2 Opinion Change Modeling with Topics

In this section we introduce the multimodal exposure feature construction used for opinion modeling. These features incorporate mobile phone sensed data as well as survey questionnaire response features. We then present the details of model selection based on a statistical approach and results for significant cases of opinion changing behavior.

6.2.1 Multimodal Exposure (MME) Features and Topics

Cumulative exposure [66], C_i to a particular political opinion O, represents the magnitude of a particular opinion that a person is exposed to on a daily basis, and is a function of the amount of contact with different individuals and their self-reported opinion. $contact_{ij}$ can be estimated from other mobile interaction features, like counts for calling, SMS, and 802.11 WLAN co-location. $contact_{ij}$ is the bluetooth proximity counts between i and j (tie-strength), and $Nbr(i)$ is the set of neighbors for i in the interaction network. C_{iO}, cumulative exposure to opinion O, from both bluetooth and call features are used for change of opinion modeling.

$$C_{iO} = \delta_j \cdot \sum_{j \in Nbr(i)} contact_{ij} \tag{6.1}$$

where $\delta_j = 1$ if person j holds opinion O, and 0 otherwise.

When considering behavioral data with Latent Dirichlet Allocation, what we refer to as *multimodal exposure (MME) features* can be seen as analogous to text words and all the data of a user is analogous to a document. Further, latent topics are analogous to human routines, where Φ in LDA gives an indication of how probable topics are for users, and Θ results in a distribution of exposure features given topics.

We formulate a multimodal vector of exposure features (MME features) encompassing four components: (1) time (2) political opinion (3) type + amount of interaction and (4) relationship. Overall, a MME feature captures the exposure to a particular political opinion, including details such as time

and relationship. Given a survey question from Table 6.1, a MME feature has the following structure $(t, p_o, b, c, f, s, p_d)$.

- Component (1) is the time where $t \in \{10$ pm - 2 am (late night = $LN)$, 2 - 8 am (early morning = $EM)$, 8 am - 5 pm (day = $D)$, 5 - 10 pm (evening = $E)\}$. These 4 time intervals in the day are specific to the overall daily activities of the users in the dataset.
- Component (2) is the political opinion $p_o \in o$ and o is the set of possible responses from Table 6.1 for the survey question chosen.
- Component (3) is the type and amount of interaction where b is a measure of the cumulative exposure from bluetooth proximity to opinion p_o and c is the cumulative exposure from the mobile phone logs to opinion p_o. Cumulative exposure, defined in Equation 6.1, is quantized into the following bins: $b \in \{0, 1-2, 2-9, 9+\}$, $c \in \{0, 1-2, 3+\}$ to limit the vocabulary size. $b = 0$ implies no proximity interaction in the time interval t with political opinion p_o and $c = 3+$ implies 3 or more calls and/or SMS with political opinion p_o during time interval t.
- Finally, the relationship metric is defined by $f \in$ [friend, not friend], $s \in$ [socialize, do not socialize], and $p_d \in$ [political discussants, not political discussants].

With this representation for users, when LDA is applied, topics are essentially clusters of dominating 'opinion exposures' present over all individuals and days in the real-life data collection, described in terms of MME features. We discuss how to determine the number of topics in the next section.

6.2.2 Model Selection

In order to choose the optimal number of topics, K, for the model, we consider statistical significance measures over the entropy of topic distributions. We chose entropy of topic distributions as it (1) enables the computation of statistical significance over a vector of probability distributions and (2) summarizes the probability distributions of user behaviors. A low entropy implies a uniform distribution over topics and a high entropy implies some topics are much more probable than others. The entropy over topic distributions is computed as follows:

$$H_i = \sum_{k=1}^{K} p(z_k|d_i) log p(z_k|d_i) \tag{6.2}$$

where d_i corresponds to the days of user i and z_k corresponds to topic k. The limitation of measure H_i is that it does not differentiate between dominating topics.

In Figure 6.3, statistical significance test results are displayed for various survey questions (Table 6.1) (e.g., interest in politics (I)) as a function of the number of topics (x-axis) for (a) the case of the two groups 'changed opinion' versus 'did not change'; and (b) considering all possible opinions and change of opinions as groups where for example for preferred party there are 4*4 possible groups. Note the baseline at $p = 0.05$ is marked in both figures with black stars to show where we consider the results to be statistically significant. The difference in group entropies is mostly statistically significant for the

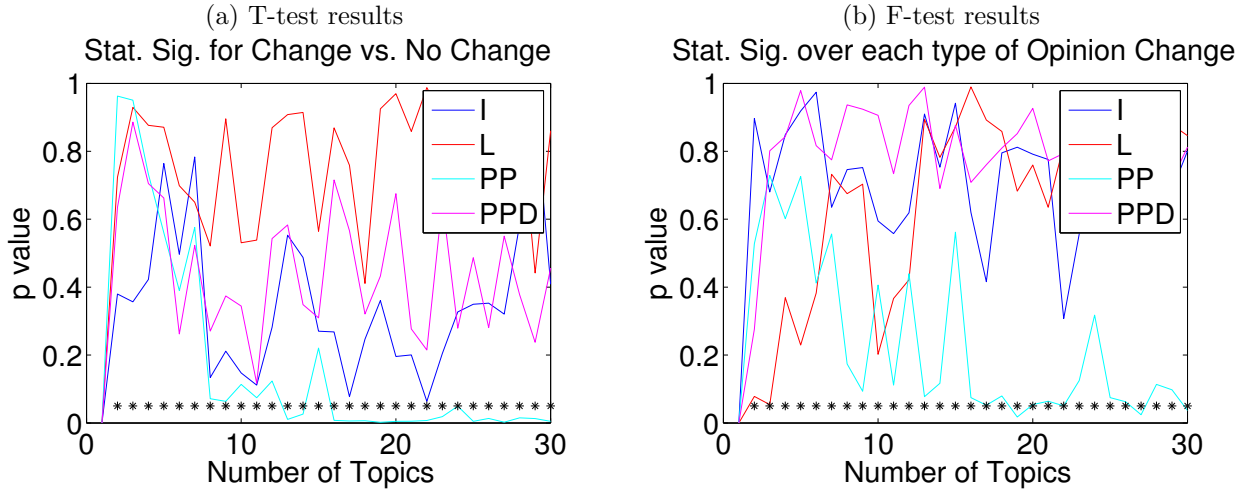


Figure 6.3. Significance results (a) for 'changed opinion' versus 'did not change opinion' for interest in politics (I), liberal/conservative (L), preferred party (PP) (4-point scale) and (PPD) (7-point scale) (b) considering all possible opinions and change of opinions as groups. The baseline in black is the level at which the p-value is considered to be statistically significant.

preferred party (PP) opinion when considering the 2 group case in (a), however not for all values of K (number of topics). In Figure 6.3 (a), the first two points for which statistical significance occurs are at $K = 13$ and $K = 14$, and in the case of Figure 6.3 (b) it occurs at $K = 17$, where $p = 0.05$ at this point. For the opinion interest in politics (I) and the 2 group case in plot (a) at $K = 22$, the p-value reaches its minimum of $p = 0.05$. We consider these points that are statistically significant in analyzing opinion change in the results.

6.2.3 Results

For experiments, we consider documents to correspond to all of the days in the life of a user. The vocabulary size of the multimodal exposure features varied depending on the opinion in question. For the opinion preferred party, it was 240 (5 possible preferred party responses * 4 possible timeslots * 4 possible cumulative exposure bins * 3 possible communication categories). For interest in politics, the vocabulary size is also 240. For both the liberal/conservative and preferred party details, the vocabulary size is 384. Experiments were run using the collapsed Gibbs sampler presented in Chapter 3 with 1000 iterations.

Interaction Patterns of People who Change Opinion. The goal is to determine the difference in the interaction patterns of the two groups people who change political opinion and people who do not. We do this by comparing the most probable topics, averaged over all the users of each of the two possible categories. By selecting the model parameters according to the points that resulted in statistical significance in Figure 6.3, and looking deeper into the differing topics or exposure patterns at these points. For the preferred party (PP), one point at which the difference in the entropy of topic distributions was statistically significant occurred at $K = 14$ topics with $p = 0.0106$. We then look at the topic distributions

of each group for $K = 14$ to see what are the difference in activities.

In Figure 6.4 (a), the top plot shows the mean Φ for those people who changed opinions and the bottom plot is for those who did not. The x-axis in both plots corresponds to the topics and the values plot are the mean $p(z|d)$ or the mean Φ averaged over all users in the group. The most probable topics (dominating routine) for users that changed opinion were topics 3, 9, and 10. The features of these three topics are visualized by Figure 6.4 (b), (c), and (d), respectively. The most dominant topic for users that did not change was topic 10, which dominated in both groups. For a given topic (Figure 6.4 (b)-(d)), we display the 3 most probable words' (top) face-to-face interaction features (middle) phone interaction features and (bottom) relationship statistics. For the top plot, or the face-to-face interaction features, the x-axis corresponds to the four time of day categories, LN=late night, EM=early morning, D=day and E=evening. The y-axis corresponds to interaction with the opinions dem=democrat, rep=republican, ind=independent, and other. The probability, nor the amount of interaction are displayed in the figure for simplification. If the value is one then one of the top words contained this feature. The middle plot contains the same characteristics as the top plot but for phone communication features (calls plus SMS). The bottom histograms indicate the relationship feature for the top words. The possible relationships, seen by the x-axis, are abbreviated by FR for friends, SOC for socialize and PD for political discussants. The colorbar identifies whether the relationship holds or not (i.e., they are friends or are not friends).

Looking at Topic 3 (Figure 6.4 (b)), we can see that users who changed opinion predominantly had face-to-face interactions with PD, who were non-friends and also not people users socialize with. The preferred party of these political discussants was democrat and this interaction occurred predominantly between 10 pm-5 pm. Further, people who changed opinion also had heavy phone call and SMS activity with democrats as well as independents. As a result, our method based on LDA was useful in determining the difference in routines between people who changed political opinion versus those that did not. Now we explore interest in politics as the opinion, to see whether or not there was differing behavior in those that increased and decreased their interest.

Different Exposure for Increased vs. Decreased Interest in Politics. We also considered the difference in daily routines of users which increased their interest in politics as opposed to those who decreased their interest. Figure 6.5 (a) shows the T-test that compares the entropy of topic distributions of both groups. Figure 6.5 (b) is the mean probability distribution of topics given the users from the two groups with $K = 22$ topics. The mean topic distribution $p(z|d)$ is shown for all users who increased their interest, and all users who decreased their interest. Plots (c)-(e) show the most probable words for the dominating topics in both groups. Topic 14 (c) is highly probable for users who increased their interest. Topic 8 and 18 are highly probable for users that decreased their interest. We disregard topics that are highly probable for both groups. By inspection, we can see that people who displayed increased interest communicated most often by phone during the day. In contrast, the group which decreased their interest had only face-to-face interactions (i.e., no phone communication) dominating their daily routines, and it included interaction with people with little and no interest as seen by topics 8 and 18. There was heavy face-to-face interactions with friends in the early morning (EM) who had no interest in politics, for the

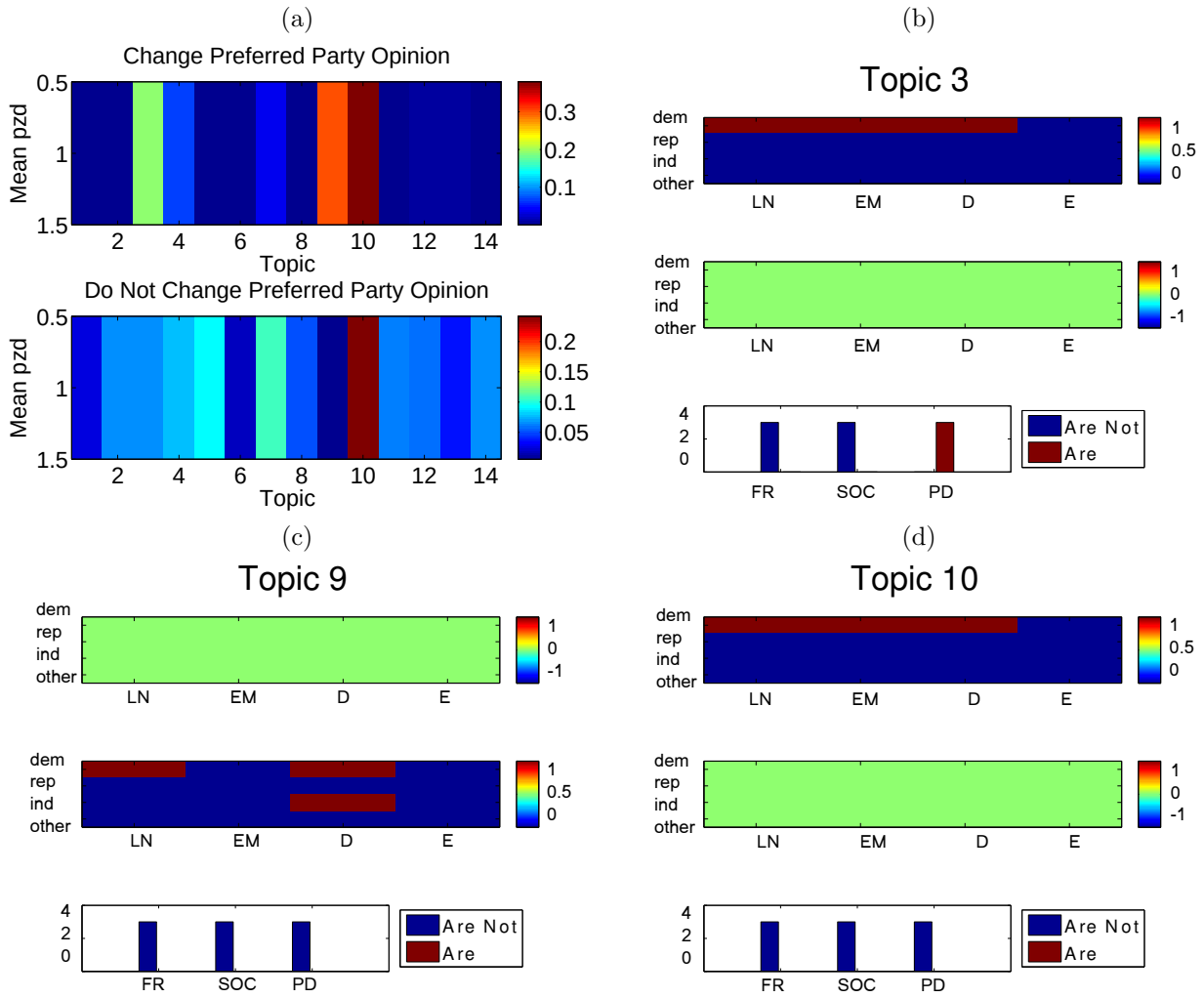


Figure 6.4. (a) Mean topic distribution of users who changed opinion (top) and users who did not (bottom). The x-axis corresponds to the topic number and the value corresponds to the mean probability of the topic over all documents. Users who changed preferred party (PP) had a high probability of topics 3, 9, 10, whereas users who did not change had a high probability of topic 10. We plot the features for topics 3, 9, and 10 in (b), (c) and (d), respectively. The top plots in (b)-(d), are the face-to-face interaction features, where the x-axis corresponds to the four time of day categories (LN=late night, EM=early morning, D=day and E=evening). The y-axis corresponds to interaction with the opinions dem=democrat, rep=republican, ind=independent, and other. The middle plot in (b)-(d) contains the same characteristics as the top plot but for phone communication features as opposed to bluetooth interaction. The bottom histograms indicate the relationship feature for the top words. The possible relationships, seen by the x-axis, are abbreviated by FR for friends, SOC for socialize and PD for political discussants. By looking at the features of the 3 most probable words for these topics, we can see that users who changed opinion displayed heavy face-to-face interactions with political discussants shown in (b), and they also had heavy phone call activity with non-friends shown in (c).

group that decreased their interest. As a result we did find some statistically significant differences in the daily routines of individuals who increased their interest in politics versus those who did not. Our experiments based on mobile phone sensor data reveal that communicating with people with little interest in politics tends to decrease one's interest in politics.

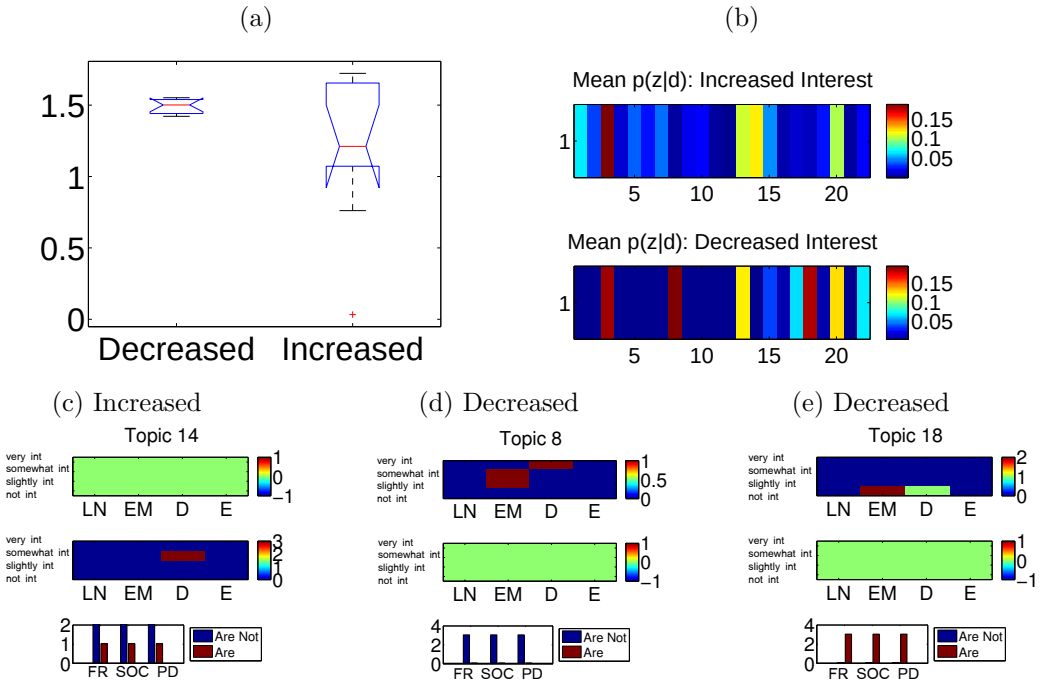


Figure 6.5. Routines of people who increased their interest in politics versus those that decreased their interest. (a) T-test results are plot for the two groups 'increased their interest in politics' and 'decreased their interest in politics' as seen by the x-axis. The y-axis corresponds to the T-test results which reveal that the difference in the entropy of topic distributions for these groups is statistically significant. (b) Mean distribution of topics for users of both groups. (c)-(e) Topics which best characterized users' daily life patterns in both groups. (c) People who increased their interest often communicated by phone. (d-e) People who decreased interest had many face-to-face interactions with people with little/no interest in politics.

6.3 Conclusion

We presented a novel application of pervasive sensing using mobile phones– modeling the spread of political opinions in real-world face-to-face networks. Using mobile phone sensors, we determined users' behaviors discovered as topics using an approach based on LDA, where features include opinions, amount of interaction, amount of phone communication as well as relationship information. We consider groups that changed opinion versus those that did not, and observed statistically significant differences in the entropy of topic distributions. This indicates that people who changed preferred party often discussed face-to-face with their democrat political discussants, and their daily routines included heavy phone and SMS activity. We also found that people who decreased their interest in politics often interacted with people who have little or no interest in politics.

One limitation of this methodology is that survey questionnaire results are noisy and can only be obtained in small scales in comparison to mobile sensor data. This limits the scope of the results and analysis. This work clearly represents a first attempt. We can anticipate several future extensions of this work. In addition to political opinions, it would be important to understand if pervasive sensing methods can help understand the propagation of other types of opinions and habits in face-to-face networks, e.g., those related to health or purchasing behavior, both in our current dataset and also in other observational data. Furthermore, with the constant improvement in sensing technologies, future projects could use global positioning system (GPS) or infra-red (IR) sensors for better location and proximity sensing. Overall, our quantitative analysis has the potential of shedding more light on long-standing open questions in political science and other social sciences, about the diffusion mechanism for opinions and behaviors, but further studies are obviously needed to realize the actual limitations of this approach.

Chapter 7

Conclusion

In this thesis we introduce methods based on probabilistic topic models to discover recurrent patterns in people's lives from socio-geographic mobile phone data. Essentially, the developed methods mine the most dominantly occurring human routines from a huge real-life corpus obtained by mobile phones to determine recurrent patterns of behavior. We investigate a number of tasks for activity modeling, mostly in an unsupervised manner, but also in a supervised manner. We developed two topic models for long duration activity modeling and apply our proposed techniques to a human-centric problem in sociology, that of opinion change.

We first investigated two different probabilistic topic models in Chapter 3, and also investigated the possible tasks in activity modeling, which can be addressed with the approach. We apply the techniques to mine daily activities of 97 mobile phone users over a 16-month period. Routines dominating the entire group's activities included "going to work late", "going home early", "working non-stop" and "having no reception" at different times. We used the routines discovered to determine behavioral patterns of users and groups of users. Furthermore, the routines discovered were used to rank users or find groups of users who display certain routines. We also characterized users based on their entropy. We compared our method to one based on clustering using k-means. Finally, we analyzed an individual's routines over time to determine regions with high variations, which may correspond to specific events.

In Chapter 4, we considered multimodal representations of location and interaction for routine discovery and proposed a method for missing data prediction based on topic models. We extended the work of Chapter 3 to formulate a socio-geographic data representation. Some of the human activities discovered with our multimodal data representation included "going out from 7 pm - midnight alone" and "working from 11 am-5 pm with 3-5 other people", further finding that this activity dominantly occurs on specific days of the week. We further demonstrated the feasibility of the topic modeling framework to predict missing multimodal phone data on specific times of the day. We also considered a supervised approach based on Support Vector Machines for single- and multimodal- cue data classification considering day type and student type.

In Chapter 5, we proposed two new models based on LDA to address existing limitations of topic

models for long duration activity modeling. The first model is the Multi-Level Topic Model (MLTM) which is essentially a multi-level LDA model, where the input of each level is formulated from the output of the previous level. The MLTM allows for the modeling of varying time duration sequences, which is a limitation of Latent Dirichlet Allocation for activity modeling. The second model is the Pairwise-Distance Topic Model, and it allows for the modeling of long duration sequences.

Finally, in Chapter 6, our approach for activity modeling based on LDA was used in the study of opinion change modeling. Specifically, we considered the differences in the daily routines of people who changed political opinions and people who did not. For this, we considered the differences in the interaction and communication patterns considering features of the individuals they are exposed to in order to differentiate between two groups in two scenarios. First we consider the difference in exposure features of individuals who changed preferred party and those who did not, as well as groups of individuals who increased their interest in politics, versus those who decreased their interest.

7.1 Limitations

While we have shown many insights into activity modeling, our work has some limitations. The first one relates to the scope of the data collection and features. The users considered in the two datasets investigated were from MIT populations, and their routines are likely not representative of society as a whole. However, the Reality Mining dataset did contain a mixture of business students as well as engineering students, and their routines are likely representative of many students and working professionals. Further, some colleagues have used our methods on the Idiap-Nokia dataset, confirming many of the dominant activities on a Swiss data collection. Another limitation is that we consider location data from cell tower connections, and reduce over 32 000 possible locations into four categories, home, work, out, and no reception. The proximity data considered is also limited in that we only consider being in proximity with other individuals from the data collection. We began some initial investigation in considering proximity to known people, strangers, laptops, and computers, however it was not pursued due to lack of ground truth. As discussed in the future work, an extension of this work would be to consider more general locations as well as interaction types. The coarse-grain timeslots considered over several hour intervals are another limitation of the methods used in several parts of this thesis. We did address this problem in Chapter 5 by introducing n-gram methods which removed the coarse-grain timeslots in our initiative to jointly model time and activity.

The second limitation of this work stems from the noise in the data. For example, what is sensed by Bluetooth proximity is often a small subset of real face-to-face interactions since most people are not carrying their mobiles in indoor settings. There is no way to account for this problem in the dataset we used and it is a limitation of all current Reality Mining studies. Other sources of noise relating to Bluetooth are the detection of other devices through thin walls where our results would reveal the two individuals are in proximity though they are not. The only way we can think of accounting for this problem is including surveys, which are highly disturbing to the individuals in the study, or to record

microphone voice features. Though the analysis of voice features would introduce new problems of its own, for example identifying the phones of the voices recorded. Other sources of noise include the noisy home and work labels provided by MIT. We chose to use the given labels for home and work, though one could try to learn these labels based on time features. The latter approach suffers in terms of missing ground truth.

A third source of limitations relates to the specific methodology chosen for this dissertation. Topic models have several limitations. There are limitations in the model parameter selection process and the objective evaluation of the routines discovered. For model selection, we use heuristic methods in many cases and we discover that overall the approach is not sensitive to model parameters. For the number of topics selected we use perplexity in some cases, and we use statistical significance measures in Chapter 6 though further insights into the optimal number of topics for each task would strengthen the results. Overall, there is a lack of ground truth for several of the tasks, many of which we have already mentioned. Further, there is no objective evaluation in the discovery of dominating activities with topics. Given more ground truth, supervised approaches could be explored and objective evaluations on unsupervised approaches would be easier to perform. However, we address the problem of evaluation by performing missing data prediction using topic models. We also compare the most probable words discovered in topics with the most frequently occurring words in the dataset to ensure we capture the critical features in our multi-level topic model approach. Finally, since the topics relate to human activities, by viewing what is mined from the data, it is apparent whether or not the methods are working. Objective evaluations are necessary however for optimal activity modeling.

7.2 Future Work

There are many directions for extending the work done in this dissertation. We discuss the possibilities relating to three general directions: the incorporation of new types of sensor data as features, the investigation of other types of machine learning models, and the development of new models for activity modeling applied to real-life scenarios.

The techniques presented in this thesis could be extended to include several types of mobile phone sensor data, including accelerometer data and mobile phone call features. Additionally, the location and Bluetooth interaction features could be explored in different ways. For example, location information can also be obtained by GPS data and WLAN information. The development of methods to encapsulate richer location features, for example cell tower identities or GPS hot spots, would reveal more detailed location information. Bluetooth devices include several categories. A device in proximity may be that of a friend or a stranger. It may also be a laptop or a computer, either owned by you or a friend or stranger. An extension of this work would be to investigate the discovery of Bluetooth devices and their types given ground truth could be collected. Knowledge of Bluetooth devices and interacting user types would be of great interest in activity discovery. In future work, the methodology for data prediction could be further optimized to use the topics in a more sophisticated manner, and to include prediction

on varying timescales, such as full days of missing data. It would also be very useful to take advantage of the other, often available data modalities of mobile sensor data for data prediction. For instance, one could predict a user's location given the time of day and their interactions, the day of the week, or even using their phone call and SMS data. The Bluetooth proximity data is potentially a very rich source if one considers proximity to all other devices including laptops, computers, and anonymous cell phones in predicting missing data. This data in itself could be used to determine the semantic labels of an individual, such as if the user is at home (in proximity with their home computer), at work (in proximity with their work computer), or out (in proximity with strangers). In a different line of work, we would like to enrich the location vocabulary by refining the "other" category. This in principle could be done from the Reality Mining dataset, but handling sparse human annotation of places is in itself a research problem. Extensions to the n-gram sequence discovery problem could be to include evaluating the models on other data types; for example, on interaction data and multimodal data. Some very promising future work relates to the modeling aspect, which we discuss next.

An obvious extension of this work would be to investigate other types of machine learning models for activity modeling. Classification methods like Support Vector Machines, and Neural Networks may be of interest for feature selection given problems containing labeled data. Hidden Markov Models and Conditional Random Fields may be good starting points for sequence modeling. Also, many of the topic models mentioned in Chapter 2, such as the Author-Recipient-Topic Model, Dynamic Topic Models, and Hierarchical Dirichlet Processes may be useful in activity modeling. Topic models can be used to address the problem of modeling the sets of interacting individuals in a meeting. Extensions of the Author-Recipient Topic Model may be of interest, where the author is the subject and the recipients are the people in proximity. However, the model would have to be modified to incorporate time dynamics and to account for the very frequently occurring case of having no one in proximity. For time considerations, the dynamic topic model could be investigated. It would be useful in studying the dynamics of topics over time. We would like to consider recent approaches like the Maximum-Margin Supervised Topic Model which explicitly addresses the issue of maximizing the distance between topics, and may be used to optimize the number of topics output. The hierarchical dirichlet processes (HDP) can also be investigated to remove the need for predefining the number of topics. One extension we plan to explore is to integrate the HDP into the pairwise-distance model. The HDP will be investigated to remove the dependency on the number of topics. Also in the pairwise-distance topic model, the j-th word depends only on the first word. In the Topical n-gram Model the j-th word depends on all of the previous words. We plan to combine both ideas with the HDP which can determine the number of previous words for dependency with the j-th word. This will likely improve the results of the Pairwise-Distance Topic Model.

Finally, new methods and mathematical tools to explore the applications of activity modeling to human-centric questions is of importance. One set of questions is in the field of epidemiology. We have access to the interaction and movements of a community. How can these be modeled and how can we use this data to address questions such as the prevention of an epidemic? Another extension relates to the analysis between communities. One research area we plan to extend is the development of a topic model

to analyze the differences and similarities between communities. To our knowledge, there is no existing topic model designed to perform such a task, which could be of great use in human-centric studies. We came across this limitation in Chapter 6 when determining the differences and similarities in behaviors of people who changed opinion versus people who did not. We overcame it by using entropy measures with statistical significance tests. However, the development of methods to quantify the differences between two communities is certainly of great interest in computational social science. There are many human-centric applications which can be explored, all requiring extensions of existing tools or new mathematical models for implementation. In terms of the political opinion modeling, there are several future extensions. In addition to political opinions, it would be important to understand if pervasive sensing methods can help understand the propagation of other types of opinions and habits in face-to-face networks, e.g., those related to health and learning.

Appendix A

Distributions and Properties

A.1 The Dirichlet Distribution

The Beta distribution is defined in Equation A.1 and the Dirichlet is defined in Equation A.2. The Dirichlet distribution generalises the Beta distribution for any dimension.

$$\begin{aligned} p(x|\alpha,\beta) &= \text{Beta}(x|\alpha,\beta) \triangleq \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \\ B(\alpha,\beta) &= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \qquad \text{where} \quad \Gamma(x+1) = x!. \end{aligned} \tag{A.1}$$

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{\alpha}) &= \text{Dirichlet}(\boldsymbol{x}|\boldsymbol{\alpha}) \triangleq \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \prod_{k=1}^{K} x_k^{\alpha_k-1} \\ &\triangleq \frac{1}{\triangle(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k-1}, \qquad \text{where} \quad \triangle(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{(dim\boldsymbol{\alpha})}\Gamma(\alpha_K)}{\Gamma(\sum_{k=1}^{(dim\boldsymbol{\alpha})}\alpha_k)} \end{aligned} \tag{A.2}$$

In this thesis, we always use a symmetric Dirichlet distribution with a scalar parameter $\alpha = \sum \alpha_k / K$ and dimension K. The Dirichlet is then defined as in Equation A.3.

$$\begin{aligned} p(\boldsymbol{x}|\alpha, K) &= \text{Dirichlet}(\boldsymbol{x}|\alpha, K) \triangleq \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{k=1}^{K} x_k^{\alpha-1} \\ &\triangleq \frac{1}{\triangle_K(\alpha)} \prod_{k=1}^{K} x_k^{\alpha-1}, \qquad \text{where} \qquad \triangle_K(\alpha) = \frac{\Gamma(\alpha)^K}{\Gamma(K\alpha)} \end{aligned} \tag{A.3}$$

A.2 The Multinomial Distribution

The multinomial distribution is defined in Equation A.4.

$$p(\boldsymbol{n}|\boldsymbol{x}, N) = \text{Multinomial}(\boldsymbol{n}|\boldsymbol{x}, N) \triangleq \frac{N!}{\prod_k n^{(k)}!} \prod_{k=1}^{K} x_k^{n^{(k)}} \tag{A.4}$$

A.3 Conjugacy

A conjugate prior, $p(\vartheta)$, of a likelihood, $p(x|\vartheta)$, is a distribution that results in a posterior distribution, $p(\vartheta|x)$, with the same functional form as the prior and a parametrisation that incorporates the observations x [42]. Conjugate prior-likelihood pairs often allow to marginalize out the likelihood parameters in closed form and thus express the likelihood of observations directly in terms of hyperparameters [42]. We use the property in Equation A.5 for the Gibbs Sampler inference procedure [42].

$$\begin{aligned} p(C|\alpha, \beta) &= \int_0^1 p(C|x)p(x|\alpha, \beta)dp \\ &= \int_0^1 x^{n^{(1)}}(1-x)^{n^{(0)}} \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}dx \\ &= \frac{1}{B(\alpha, \beta)} \int_0^1 x^{n^{(1)}+\alpha-1}(1-x)^{n^{(0)}+\beta-1}dx \\ &= \frac{B(n^{(1)}+\alpha, n^{(0)}+\beta)}{B(\alpha, \beta)} \\ &= \frac{\Gamma(n^{(0)}+\beta)\Gamma(n^{(1)}+\alpha)}{\Gamma(n^{(0)}+n^{(1)}+\alpha+\beta)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \end{aligned} \tag{A.5}$$

Appendix B

Pairwise-Distance Topic Model Inference Derivation Details

$$
\begin{aligned}
p(\boldsymbol{z}|\alpha) &= \int p(\boldsymbol{z}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\alpha)d\boldsymbol{\theta} \\
p(\boldsymbol{z}|\boldsymbol{\theta}) &= \prod_{m=1}^{M}\prod_{k=1}^{T}\boldsymbol{\theta}_{m,k}^{n_m^k} \\
p(\boldsymbol{\theta}|\alpha) &= \prod_{m=1}^{M}\frac{1}{B(\alpha)}\prod_{k=1}^{T}\boldsymbol{\theta}_{m,k}^{\alpha-1}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
p(\boldsymbol{z}|\alpha) &= \prod_{m=1}^{M}(\frac{1}{B(\alpha)}\int\prod_{k=1}^{T}\boldsymbol{\theta}_{m,k}^{n_m^k+\alpha-1}d\boldsymbol{\theta}) \\
&= \prod_{m=1}^{M}\frac{B(n_m+\alpha)}{B(\alpha)} \qquad \text{where} \quad n_m = \{n_m^k\}_{k=1}^{T}.
\end{aligned}
\tag{B.1}
$$

Similarly,

$$\begin{aligned} p(\boldsymbol{w_1}|\boldsymbol{z},\beta_1) &= \int p(\boldsymbol{w_1}|\boldsymbol{z},\phi_1)p(\phi_1|\beta_1)d\phi_1 \\ p(\boldsymbol{w_1}|\boldsymbol{z},\phi_1) &= \prod_{k=1}^{T}\prod_{t=1}^{V}\phi_{1_{k,t}}^{n_k^t} \\ p(\phi_1|\beta_1) &= \prod_{k=1}^{T}\frac{1}{B(\beta_1)}\prod_{t=1}^{V}\phi_{1_{k,t}}^{\beta_1-1} \end{aligned}$$

Therefore,

$$\begin{aligned} p(\boldsymbol{w_1}|\boldsymbol{z},\beta_1) &= \prod_{k=1}^{T}(\frac{1}{B(\beta_1)}\int\prod_{t=1}^{V}\phi_{1_{k,t}}^{n_k^t+\beta_1-1}d\phi_1) \\ &= \prod_{k=1}^{T}\frac{B(n_k+\beta_1)}{B(\beta_1)} \qquad \text{where} \quad n_k = \{n_k^t\}_{t=1}^{V}. \end{aligned} \tag{B.2}$$

Considering all j for which $1 < j \leqq n$,

$$\begin{aligned} p(\boldsymbol{w_j}|\boldsymbol{w_1},z,\beta_j) &= \int p(\boldsymbol{w_j}|\boldsymbol{w_1},\boldsymbol{z},\phi_j)p(\phi_j|\beta_j)d\phi_j \\ p(\boldsymbol{w_j}|\boldsymbol{w_1},\boldsymbol{z},\phi_j) &= \prod_{k=1}^{T}\prod_{t_1=1}^{V}\prod_{t_2=1}^{V}\phi_{j_{k,t_1,t_2}}^{n_{k_j^{`}}^{(t_1,t_2)_j}} \\ p(\phi_j|\beta_j) &= \prod_{k=1}^{T}\frac{1}{B(\beta_j)}\prod_{t_1=1}^{V}\prod_{t_2=1}^{V}\phi_{j_{k,t_1,t_2}}^{\beta_j-1} \end{aligned}$$

Therefore,

$$\begin{aligned} p(\boldsymbol{w_j}|\boldsymbol{w_1},\boldsymbol{z},\beta_j) &= \prod_{k=1}^{T}\frac{1}{B(\beta_j)}(\int\prod_{t_1=1}^{V}\prod_{t_2=1}^{V}\phi_{j_{k,t_1,t_2}}^{n_{k_i^{`}}^{(t_1,t_2)_j}+\beta_j-1}d\phi_j) \\ &= \prod_{k=1}^{T}\frac{B(n_{k_j^{`}}+\beta_j)}{B(\beta_j)} \qquad \text{where} \quad n_{k_j^{`}} = \{n_{k_j^{`}}^{(t_1,t_2)_j}\}_{t_1=1,t_2=1}^{V,V} \end{aligned} \tag{B.3}$$

$$p(z_i = k|\mathbf{z}_{-i}, \mathbf{q}, \alpha, \beta) = \frac{p(\mathbf{z}, \mathbf{q}|\alpha, \beta)}{p(\mathbf{z}_{-i}, \mathbf{q}|\alpha, \beta)} \tag{B.4}$$

$$= \frac{p(\mathbf{z}|\alpha)}{p(\mathbf{z}_{-i}|\alpha)} \cdot \frac{p(\boldsymbol{w_1}|\mathbf{z}, \beta_1)}{p(\boldsymbol{w_{1}}_{-i}|\mathbf{z}_{-i}, \beta_1) \cdot p(\boldsymbol{w_{1}}_{i})} \cdot \prod_{j=2}^{n} \frac{p(\boldsymbol{w_j}|\boldsymbol{w_1}, \mathbf{z}, \beta_j)}{p(\boldsymbol{w_j}_{-i}|\mathbf{z}_{-i}, \boldsymbol{w_{1}}_{-i}, \beta_j) \cdot p(\boldsymbol{w_j}_{i})} \tag{B.5}$$

$$\propto \frac{B(n_m + \alpha)}{B(n_{m_{-i}} + \alpha)} \cdot \frac{B(n_k + \beta_1)}{B(n_{k_{-i}} + \beta_1)} \cdot \prod_{j=2}^{n} \frac{B(n_{k'_j} + \beta_j)}{B(n_{k'_{j,-i}} + \beta_j)} \tag{B.6}$$

(B.7)

$$= \frac{\Gamma(n_m^k + \alpha)\Gamma(\sum_{k=1}^{K} n_{m,-i}^k + \alpha)}{\Gamma(n_{m,-i}^k + \alpha)\Gamma(\sum_{k=1}^{K} n_m^k + \alpha)} \cdot \frac{\Gamma(n_k^t + \beta_1)\Gamma(\sum_{t=1}^{V} n_{k,-i}^t + \beta_1)}{\Gamma(n_{k,-i}^t + \beta_1)\Gamma(\sum_{t=1}^{V} n_k^t + \beta_1)} \cdot \prod_{j=2}^{n} \frac{\Gamma(n_k^t + \beta_j)\Gamma(\sum_{t_1=1}^{V}\sum_{t_2=1}^{V} n_{k,-i}^{(t_1,t_2)_j} + \beta_j)}{\Gamma(n_{k,-i}^{(t_1,t_2)_j} + \beta_j)\Gamma(\sum_{t_1=1}^{V}\sum_{t_2=1}^{V} n_k^{(t_1,t_2)_j} + \beta_j)} \tag{B.8}$$

$$= \frac{n_m^k + \alpha - 1}{\sum_k n_m^k + \alpha - 1} \cdot \frac{n_k^t + \beta_1 - 1}{\sum_{t=1}^{V} n_k^t + \beta_1 - 1} \cdot \prod_{j=2}^{n} \frac{n_k^{(t_1,t_2)_j} + \beta_j - 1}{\sum_{t_1=1}^{V}\sum_{t_2=1}^{V} n_k^{(t_1,t_2)_j} + \beta_j - 1} \tag{B.9}$$

$$\propto (n_{m,-i}^k + \alpha) \cdot \frac{n_{k,-i}^t + \beta_1}{\sum_{t=1}^{V} n_{k,-i}^t + \beta_1} \cdot \prod_{j=2}^{n} \frac{n_{k,-i}^{(t_1,t_2)_j} + \beta_j}{\sum_{t_1=1}^{V}\sum_{t_2=1}^{V} n_{k,-i}^{(t_1,t_2)_j} + \beta_j} \tag{B.10}$$

since $n_{m,-i}^k = n_m^k - 1$. Also $\sum_k n_m^k + \alpha$ is always constant for any potential $z_i = k$ therefore it can be ignored. $n_k = \{n_k^{(t)}\}_{t=1}^{V}$ and $n_{k'_j} = \{n_{k'}^{(t_1,t_2)_j}\}_{t_1=1,t_2=1}^{t_1=V,t_2=V}$. We use the properties $B(x) = \frac{\prod_{k=1}^{dim x}\Gamma(x_k)}{\Gamma(\sum_{k=1}^{dim x} x_k)}$, and $\Gamma(y) = (y-1)!$.

Bibliography

[1] Akoush, S. and Sameh, A. (2007). Mobile user movement prediction using bayesian learning for neural networks. In *International Wireless Communications and Mobile Computing Conference (ACM IWCMC)*, pages 191–196, Honolulu, Hawaii, USA.

[2] Asuncion, A., Welling, M., Smyth, P., and Teh, Y. W. (2009). On smoothing and inference for topic models. In *Proc. of the Conference on Uncertainty in Artificial Intelligence (UAI)*, Montreal, Canada.

[3] Bernard, H., Killworth, P., Kronenfeld, D., and Sailer, L. (1984). The problem of informant accuracy: The validity of retrospective data. *Annual Reviews in Anthropology*.

[4] Bimbot, F., Pieraccini, R., Levin, E., and Atal, B. (1995). Variable-length sequence modeling: multigrams. *IEEE Signal Processing Letters*, **2**(6), 111–113.

[5] Bishop, C. M. (2007). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1st ed. 2006. corr. 2nd printing edition.

[6] Blei, D. and Lafferty, J. (2006). Dynamic topic models. In *Proc. of the 23rd International Conference on Machine Learning (ICML)*, Pittsburgh, Pennsylvania, USA.

[7] Blei, D. and Lafferty, J. (2007). A correlated topic model of science. *Annals of Applied Statistics*, **1**(1), 17–35.

[8] Blei, D. and Lafferty, J. (2009). Topic models. In A. Srivastava and M. Sahami, editors, *Text Mining: Theory and Applications*. Taylor and Francis.

[9] Blei, D. and McAuliffe, J. (2007). Supervised topic models. In *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, Canada.

[10] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**, 993–1022.

[11] Blei, D. M., Griffiths, T. L., Jordan, M. I., and Tenenbaum, J. B. (2004). Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, Canada.

[12] Brewer, D. and Webster, C. (2000). Forgetting of friends and its effects on measuring friendship networks. *Social Networks*, **21**(4), 361 – 373.

[13] Candia, J., Gonzalez, M., Wang, P., Schoenharl, T., Madey, G., and Barabasi, A.-L. (2008). Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, **41**(22), 224015+.

[14] Choudhury, T. (2003). *Sensing and Modeling Human Networks*. Ph.D. thesis, M.I.T.

[15] Choudhury, T. and Basu, S. (2004). Modeling conversational dynamics as a mixed memory markov process. In *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, Canada.

[16] Choudhury, T. and Pentland, A. (2003). Sensing and modeling human networks using the sociometer. In *IEEE International Symposium on Wearable Computers (ISWC)*, pages 216–, Washington, USA.

[17] Choudhury, T., Philipose, M., Wyatt, D., and Lester, J. (2006). Towards activity databases: Using sensors and statistical models to summarize people's lives. In *IEEE Data Eng. Bull*, pages 49–58.

[18] Chronis, L., Madan, A., and Pentland, A. (2009). Socialcircuits: The art of using mobile phones for modeling personal interactions. In *ICMI-MLMI Workshop on Multimodal Sensor-Based Systems and Mobile Phones for Social Computing*, Cambridge, Massachusetts, USA.

[19] Church, K. and Gale, W. (1991). Concordances for parallel text. In *Proc. of the 7th Annual Conf. of the UW Centre for the New OED and Text Research*, pages 40–62.

[20] Davis, M., Van House, N., Towle, J., King, S., Ahern, S., Burgener, C., Perkel, D., Finn, M., Viswanathan, V., and Rothenberg, M. (2005). MMM2: mobile media metadata for media sharing. In *Extended Abstracts on Human Factors in Computing Systems (CHI)*, pages 1335–1338, Portland, Oregan, USA.

[21] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the Amercian Society for Information Science*, **41**(6), 391–407.

[22] Duda, R., Hart, P., and Stork, D. (2000). *Pattern Classification (2nd Ed.)*. Wiley-Interscience.

[23] Dunning, T. E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1), 61–74.

[24] Duong, T. V., Bui, H. H., Phung, D. Q., and Venkatesh, S. (2005). Activity recognition and abnormality detection with the switching hidden semi-markov model. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) - Volume 1*, pages 838–845, Washington, USA.

[25] Eagle, N. and Pentland, A. (2009). Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, **63**(7), 1057–1066.

[26] Eagle, N., Pentland, A., and Lazer, D. (2009). Inferring social network structure using mobile phone data. *Proc. of the National Academy of Sciences (PNAS)*, **106**(36), 15274–15278.

[27] Eagle, N., Macy, M., and Claxton, R. (2010). Network diversity and economic development. *Science*, **328**(5981), 1029–1031.

[28] Erosheva, E., Fienberg, S., and Lafferty, J. (2004). Mixed-membership models of scientific publications. *Proc. of the National Academy of Sciences (PNAS USA)*, **101**(Suppl 1), 5220–5227.

[29] Farrahi, K. and Gatica-Perez, D. (2008a). Daily routine classification from mobile phone data. In *5th Joint Workshop on Machine Learning and Multimodal Interaction (MLMI)*, Utrecht, Netherlands.

[30] Farrahi, K. and Gatica-Perez, D. (2008b). Discovering human routines from cell phone data with topic models. In *IEEE International Symposium on Wearable Computers (ISWC)*, Pittsburgh, Pennsylvania, USA.

[31] Farrahi, K. and Gatica-Perez, D. (2008c). What did you do today? Discovering daily routines from large-scale mobile data. In *ACM Int. Conf. on Multimedia (ACM MM)*, Vancouver, Canada.

[32] Farrahi, K. and Gatica-Perez, D. (2009). Learning and predicting multimodal daily life patterns from cell phones. In *ICMI-MLMI*, Cambridge, Massachusetts USA.

[33] Farrahi, K. and Gatica-Perez, D. (2010a). Mining human location-routines using a multi-level topic model. In *Socialcom Symposium on Social Intelligence and Networking (Socialcom SIN-10)*.

[34] Farrahi, K. and Gatica-Perez, D. (2010b). Probabilistic mining of socio-geographic routines from mobile phone data. *IEEE Journal of Selected Topics in Signal Processing (J-STSP)*, **4**(4), 746–755.

[35] Farrahi, K. and Gatica-Perez, D. (2011). Discovering routines from large-scale human locations using probabilistic topic models. *ACM Transactions on Intelligent Systems and Technology, Special Issue on Intelligent Systems for Activity Recognition*, **2**(1), 3:1–3:27.

[36] Fay, N., Garrod, S., and Carletta, J. (2000). Group discussion as interactive dialogue or as serial monologue: The influence of group size. *Psychological Science*, **English**, 481–486+.

[37] Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, **453**(7196), 779–782.

[38] Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proc. of the National Academy of Sciences (PNAS USA)*, **101 Suppl 1**, 5228–5235.

[39] Guo, F., Hanneke, S., Fu, W., and Xing, E. P. (2007). Recovering temporally rewiring networks: A model-based approach. In *Proc. of the International Conference on Machine Learning (ICML)*, Oregan, USA.

[40] Hamid, R., Maddi, S., Johnson, A., Bobick, A., Essa, I., and Isbell, C. (2005). Unsupervised activity discovery and characterization from event-streams. In *Proc. of the Conference on Uncertainty in Artificial Intelligence (UAI)*, Edinburgh, Scotland.

[41] Hamid, R., Maddi, S., Bobick, A. F., and Essa, I. A. (2007). Structure from statistics - unsupervised activity analysis using suffix trees. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–8, Rio de Janeiro, Brazil.

[42] Heinrich, G. (2008). Parameter estimation for text analysis,. Technical report, University of Leipzig.

[43] Hidalgo, C. A. and Rodriguez-Sickert, C. (2008). The dynamics of a mobile phone network. *Physica A*, **387**, 3017–3024.

[44] Hightower, J., Consolvo, S., Lamarca, A., Smith, I., and Hughes, J. (2005). Learning and recognizing the places we go. In *Ubiquitous Computing (UbiComp)*, pages 159–176, Tokyo, Japan.

[45] Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proc. of the Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 289–296, Stockholm, Sweden.

[46] Huckfeldt, R. and Sprague, J. (1991). Discussant Effects on Vote Choice: Intimacy, Structure and Interdependence. *The Journal of Politics*, **53**, 122–158.

[47] Hunter, D. R. and Handcock, M. S. (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, **15**, 565–583.

[48] Huynh, T., Fritz, M., and Schiele, B. (2008). Discovery of activity patterns using topic models. In *Ubiquitous computing (UbiComp)*, pages 10–19, Seoul, Korea.

[49] Intille, S. S., Larson, K., Tapia, E. M., Beaudin, J. S., Kaushik, P., Nawyn, J., and Rockinson, R. (2006). Using a live-in laboratory for ubiquitous computing research. In *Pervasive Computing*, pages 349–365.

[50] Jebara, T. (2008). http://www.sensenetworks.com. Sense Networks.

[51] Justeson, J. S. and Katz, S. M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, **1**.

[52] Kiukkonen, N., Blom, J., Dousse, O., Gatica-Perez, D., and Laurila, J. (2010). Towards rich mobile phone datasets: Lausanne data collection campaign. In *Proc. ACM Int. Conf. on Pervasive Services (ICPS)*, Berlin, Germany.

[53] Krumm, J. and Horvitz, E. (2006). Predestination: Inferring destinations from partial trajectories. In *Ubiquitous Computing (UbiComp)*, California, USA.

[54] Laerhoven, K. V., Borazio, M., Kilian, D., and Schiele, B. (2008). Sustained logging and discrimination of sleep postures with low-level, wrist-worn sensors. In *IEEE International Symposium on Wearable Computers (ISWC)*, pages 69–76, Pittsburgh, Pennsylvania, USA.

[55] Larson, K. and Intille, S. (2002). http://architecture.mit.edu/house_n/data/placelab/placelab.htm. Placelab website.

[56] Lazer, D., Rubineau, B., Katz, N., Chetkovich, C., and Neblo, M. A. (2008). Networks and political attitudes: Structure, influence, and co-evolution. Working paper series, Harvard University.

[57] Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Van Alstyne, M. (2009). Computational social science. *Science*, **323**(5915), 721–723.

[58] Letchner, J., Fox, D., and LaMarca, A. (2005). Large-scale localization from wireless signal strength. In *National Conference on Artificial Intelligence (AAAI)*, pages 15–20, Pittsburgh, Pennsylvania, USA.

[59] Liao, L., Fox, D., and Kautz, H. (2006). Location-based activity recognition. In *Advances in Neural Information Processing Systems (NIPS)*, pages 787–794, Vancouver, Canada.

[60] Loecher, M. and Jebara, T. (2009). CitySense: Multiscale space time clustering of gps points and trajectories. In *Proc. of the Joint Statistical Meeting*.

[61] Logan, B., Healey, J., Philipose, M., Tapia, E. M., and Intille, S. S. (2007). A long-term evaluation of sensing modalities for activity recognition. In *Ubiquitous Computing (UbiComp)*, volume 4717 of *Lecture Notes in Computer Science*, pages 483–500. Springer.

[62] Lu, H., Pan, W., Lane, N. D., Choudhury, T., and Campbell, A. T. (2009). Soundsense: scalable sound sensing for people-centric applications on mobile phones. In *Proc. of the 7th International Conference on Mobile Systems, Applications, and Services (MobiSys)*, pages 165–178, New York, NY, USA.

[63] Madan, A. and Pentland, A. (2006). Vibefones: Socially aware mobile phones. In *IEEE International Symposium on Wearable Computers (ISWC)*, Montreux, Switzerland.

[64] Madan, A., Cebrián, M., Lazer, D., and Pentland, A. (2010a). Social sensing for epidemiological behavior change. In *Ubiquitous Computing (UbiComp)*, pages 291–300, Copenhagen, Denmark.

[65] Madan, A., Moturu, S., Lazer, D., and Pentland, A. (2010b). Social sensing: Obesity, unhealthy eating and exercise in face-to-face networks. In *Proc. of ACM Wireless Health*, San Diego, USA.

[66] Madan, A., Farrahi, K., Gatica-Perez, D., and Pentland, A. (2011). Pervasive sensing to model political opinions in face-to-face networks. In *Pervasive*, San Francisco, USA.

[67] Mccallum, A., Corrada-emmanuel, A., and Wang, X. (2004). The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email. Technical report, University of Massachusetts Amherst.

[68] McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., and Zhang, D. (2005). Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, **27**(3), 305–317.

[69] Miluzzo, E., Lane, N. D., Eisenman, S. B., and Campbell, A. T. (2007). Cenceme: injecting sensing presence into social networking applications. In *Proc. of the 2nd European conference on Smart sensing and context*, EuroSSC'07, pages 1–28, Berlin, Heidelberg. Springer-Verlag.

[70] Monay, F. and Gatica-Perez, D. (2007). Modeling semantic aspects for cross-media image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, **29**(10), 1802–1817.

[71] Nallapati, R. and Cohen, W. (2008). Link-plsa-lda: A new unsupervised model for topics and influence in blogs. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, Seattle, Washington, USA.

[72] Niebles, J., Wang, H., and Fei-Fei, L. (2008). Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision (IJCV)*, **79**, 299–318.

[73] Olguin, D., Waber, B., Kim, T., Mohan, A., Ara, K., and Pentland, A. (2009). Sensible organizations: Technology and methodology for automatically measuring organizational behavior. *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, **39**(1), 43–55.

[74] Olguin Olguin, D., Gloor, P., and Pentland, A. (2009). Wearable sensors for pervasive healthcare management. *PCT Healthcare*.

[75] Otsason, V., Varshavsky, A., Lamarca, A., and de Lara, E. (2005). Accurate gsm indoor localization. In *Ubiquitous Computing (UbiComp)*, pages 141–158. Springer Berlin / Heidelberg.

[76] Patel, S., Kientz, J., Hayes, G., Bhat, S., and Abowd, G. (2006). Farther than you may think: An empirical investigation of the proximity of users to their mobile phones. In *Proc. of Ubicomp*, pages 123–140, Orange County, California, USA.

[77] Patterson, D., Liao, L., Fox, D., and Kautz, H. (2003). Inferring high-level behavior from low-level sensors. In *Ubiquitous Computing (UbiComp)*, pages 73–89, Seattle, USA.

[78] Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, **155**(2), 945–959.

[79] Quelhas, P., Monay, F., Odobez, J.-M., Gatica-Perez, D., and Tuytelaars, T. (2007). A thousand words in a scene. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, **29**(9), 1575–89.

[80] Reddy, S., Burke, J., Estrin, D., Hansen, M., and Strivastava, M. (2008). Using mobile phones to determine transportation mode. In *IEEE International Symposium on Wearable Computers (ISWC)*, Pittsburgh, Pennsylvania, USA.

[81] Robins, G., Snijders, T., Wang, P., Handcock, M., and Pattison, P. (2007). Recent developments in exponential random graph (p*) models for social networks. *Social Networks*, **29**(2), 192–215.

[82] Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. (2004). The author-topic model for authors and documents. In *Proc. of the Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 487–494, Arlington, USA.

[83] Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, **19**.

[84] Sohn, T., Varshavsky, A., Lamarca, A., Chen, M., Choudhury, T., Smith, I., Consolvo, S., Hightower, J., Griswold, W., and de Lara, E. (2006). Mobility detection using everyday gsm traces. In *Ubiquitous Computing (UbiComp)*, pages 212–224, California, USA.

[85] Steyvers, M., Smyth, P., Rosen-Zvi, M., and Griffiths, T. (2004). Probabilistic author-topic models for information discovery. In *International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD)*, pages 306–315, New York, USA.

[86] Tapia, E. M., Intille, S. S., and Larson, K. (2004). Activity recognition in the home setting using simple and ubiquitous sensors. In A. Ferscha and F. Mattern, editors, *Pervasive Computing*, volume 3001 of *Lecture Notes in Computer Science*, pages 158–175. Springer Berlin / Heidelberg.

[87] Tapia, E. M., Intille, S. S., Haskell, W., Larson, K., Wright, J., King, A., and Friedman, R. (2007). Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart monitor. In *IEEE International Symposium on Wearable Computers (ISWC)*, Boston, USA.

[88] Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, **101**(476), 1566–1581.

[89] Teh, Y. W., Newman, D., and Welling, M. (2007). A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, Canada.

[90] Wallach, H. (2006). Topic modeling: beyond bag-of-words. In *Proc. of the International Conference on Machine Learning (ICML)*, Pittsburgh, Pennsylvania, USA.

[91] Wang, P., Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L. (2009a). Understanding the spreading patterns of mobile phone viruses. *Science*, **324**(5930), 1071–1076.

[92] Wang, X., Mohanty, N., and Mccallum, A. (2005). Group and topic discovery from relations and text. In *International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD) Workshop on Link Discovery*, Chicago, USA.

[93] Wang, X., McCallum, A., and Wei, X. (2007). Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *IEEE International Conference on Data Mining (ICDM)*, pages 697–702, Washington, USA.

[94] Wang, X., Ma, X., and Grimson, W. E. L. (2009b). Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, **31**(3), 539–555.

[95] Wesolowski, A. and Eagle, N. (2010). Parameterizing the dynamics of slums. In *AAAI Spring Symposium on Artificial Intelligence for Development (AI-D)*.

[96] Wojek, C., Nickel, K., and Stiefelhagen, R. (2006). Activity recognition and room-level tracking in an office environment. In *IEEE International Conf. on Multisensor Fusion and Integration for Intelligent Systems*, pages 25–30.

[97] Wren, C., Ivanov, Y., Kaur, I., Leigh, D., Westhues, J., Wren, C. R., Ivanov, Y. A., Kaur, I., Leigh, D., and Westhues, J. (2007). Socialmotion: Measuring the hidden social life of a building. In *International Symposium on Location- and Context-Awareness (LoCA 2007)*, pages 85–102.

[98] Wyatt, D., Choudhury, T., and Kautz, H. (2007). Capturing spontaneous conversation and social dynamics: A privacy sensitive data collection effort. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Honolulu, Hawaii, USA.

[99] Wyatt, D., Choudhury, T., and Bilmes, J. (2010). Discovering long range properties of social networks with multi-valued time-inhomogeneous models. In *National Conference on Artificial Intelligence (AAAI)*, Atlanta, Georgia, USA.

[100] Yano, T., Cohen, W. W., and Smith, N. A. (2009). Predicting response to political blog posts with topic models. In *North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT)*, pages 477–485.

Curriculum Vitae

Idiap Research Institute
Rue Marconi 19
PO Box 592
CH-1920 Martigny
Switzerland
☎ *+41 (0)27 721 77 97*
✉ *katayounfarrahi@gmail.com*
www.idiap.ch/∼kfarrahi/

Katayoun Farrahi

Canadian citizen. Swiss resident.

Research Interests

Reality Mining, machine learning, topic models, activity modeling, computational social science.

Professional Experience

February 2007 – present
Research Assistant, *Idiap Research Institute*, Martigny, Switzerland.
Conducting research on human mobile phone data collections, called Reality Mining. Investigating machine learning tools, specifically topic models, for the unsupervised discovery and analysis of large-scale human activity patterns. *http://www.idiap.ch/∼kfarrahi*

July 2006 – January 2007
Software Engineer, *Broadcom Corporation*, Vancouver B.C., Canada.
Worked with key customers to find and resolve technical issues in our embedded software application for the Linux operating system. Adjusted the Phonexchange software to support multiple simultaneous custom ring tones. *http://www.broadcom.com/*

February 2005 – July 2006
Multimedia Software developer, *Counterpath Solutions*, Vancouver B.C., Canada.
Implemented the API for the compression of audio in our VoIP SDK. Incorporated an H.264 video codec into our application. Implemented an RFC for the packetization of AMR-WB for RTP. Developed a test framework which runs nightly and evaluates the audio and video quality. Worked with key customers to solve a wide range of problems in audio, video, and networking protocols. *http://www.counterpath.com/*

June 2002 – August 2004
Research Assistant, *University of Victoria*, Victoria B.C., Canada.
Reduced RTP packet loss of H.264/AVC video over 3G WCDMA Networks using Forward Error Correction and Interleaving. *http://www.ece.uvic.ca/∼ agullive/*

Awards

2010 Idiap PhD Student Research Award.

2010 2010 Google Anita Borg Memorial Scholarship.

2009 ICMI conference travel grant, from the National Science Foundation (NSF). (1200 USD)

2003 Natural Sciences and Engineering Research Council of Canada (NSERC) Industrial Post-Graduate Scholarship.

2002 University of Victoria President's Scholarship.

2002 University of Victoria Research Assistantship.

Software

OS	Linux, Unix, Windows, Mac	Other	LaTeX
programming	C/C++, Python	database	MySQL
scientific	Matlab	web development	html

Education

2007–present **Ph.D. in Computer Science**, *École Polytechnique Fédérale de Lausanne (EPFL) and Idiap Research Institute, Martigny*.
Advisor: Dr. Daniel Gatica-Perez
Thesis: A Probabilistic Approach to Socio-Geographic Reality Mining (See Doctoral Dissertation section below for details.)

2002–2004 **Master of Electrical Engineering, Wireless Communications**, *University of Victoria*.
Advisor: Prof. Aaron Gulliver
Thesis: Robust H.264/AVC Video Transmission over 3G WCDMA Networks
Investigated error resilient and concealment methods for robust wireless video transmission.

1998–2002 **Bachelor of Engineering Science, Electrical Option**, *University of Toronto*.
Advisor: Prof. Ali Sheikholeslami
Thesis: Phase Interpolator Circuits for High Speed Clock Generation
Implemented and studied a phase interpolator circuit for high speed clock generation.

Doctoral Dissertation

Title: A Probabilistic Approach to Socio-Geographic Reality Mining

Description: Investigated machine learning approaches, particulary probabilistic topic models, for large-scale human activity discovery. Developed methods based on Latent Dirichlet Allocation for socio-geographic (large-scale location and interaction) activity mining. Developed the Iterative Topic Model, an extension of LDA, for varying-duration sequence discovery. Also developed the Pairwise-Distance Topic Model for long-duration sequence discovery. Applied developed approaches to a computational social science problem, that of human political opinion change.

Programming languages used: Matlab, C/C++, MySQL, html, LaTeX.

Publications - Journals

[1] Probabilistic Mining of Socio-Geographic Routines from Mobile Phone Data,
K. Farrahi and D. Gatica-Perez, *IEEE Journal of Special Topics in Signal Processing, Vol. 4, No. 4, pp.746-755*, Aug. 2010.

[2] Discovering Routines from Large-Scale Human Locations using Hierarchical Bayesian Models, K. Farrahi and D. Gatica-Perez, *ACM Transactions on Intelligent Systems and Technology, Special Issue on Intelligent Systems for Activity Recognition, Vol. 2, No. 1*, Jan. 2011.

Publications - Conferences

[1] Pervasive Sensing to Model Political Opinions in Face-to-Face Networks,
A. Madan, K. Farrahi, D. Gatica-Perez, and A. Pentland, *Pervasive*, San Francisco, USA, June 2011.

[2] Mining Human Location-Routines using a Multi-Level Topic Model,
K. Farrahi and D. Gatica-Perez, *Socialcom Symposium on Social Intelligence and Networking (Socialcom SIN-10)*, Minneapolis, USA, Aug. 2010.

[3] Learning and Predicting Multimodal Daily Life Patterns from Cell Phones,
K. Farrahi and D. Gatica-Perez, *ICMI-MLMI*, Cambridge, USA, Nov. 2009. (Accepted for Oral Presentation as well as Doctoral Symposium poster) .

[4] Discovering Human Routines from Cell Phone Data with Topic Models,
K. Farrahi and D. Gatica-Perez, *12th IEEE International Symposium on Wearable Computers (ISWC)*, Pittsburgh, USA, 2008. (Accepted for Oral Presentation)

[5] What Did You Do Today? Discovering Daily Routines from Large-Scale Mobile Data, K. Farrahi and D. Gatica-Perez, *ACM International Conference on Multimedia (ACM MM)*, Vancouver, Canada, 2008.

[6] Daily Routine Classification from Mobile Phone Data, K. Farrahi and D. Gatica-Perez, *5th Joint Workshop on Machine Learning and Multimodal Interaction (MLMI)*, Utrecht, Netherlands, 2008. (Accepted for Oral Presentation)

[7] Robust Video Transmission using Forward Error Correction over 3G WCDMA Networks, K. Farrahi and T.A. Gulliver, *IEEE WoWCaS*, Vancouver, Canada, May 2004. (Accepted for Oral Presentation)

Projects

April – June 2010 **Internship**, *Massachusetts Institute of Technology (MIT)*.
Visiting The Human Dynamics Lab led by Prof. Alex Pentland at the MIT Media Lab. Understanding the underlying behaviors in human opinion change from Reality Mining data.

1999 – 2000 **Robot Design and Manufacture**, *University of Toronto*.
Designed and built a successful computer-controlled pancake maker. Responsible for the mechanical component of the robot. Worked closely with the person in charge of the electrical component.

2001 – 2002 **4th Year Undergraduate Thesis Project**, *University of Toronto*.
Implemented and analyzed a phase interpolator circuit for high-speed multiphase clock generation.

Languages

English *(native)*, French *(intermediate)*, Persian *(advanced)*