LOCATION-BASED SOCIAL NETWORKS: LATENT TOPICS MINING AND HYBRID TRUST-BASED RECOMMENDATION

by

Xuelian Long

B.S. in Electrical Engineering, Chongqing University of Posts and Telecommunications, 2005

M.E. in Electronic and Communication Engineering, Beijing University of Posts and Telecommunications, 2008

Submitted to the Graduate Faculty of

the School of Information Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2015

UNIVERSITY OF PITTSBURGH

SCHOOL OF INFORMATION SCIENCES

This dissertation was presented

by

Xuelian Long

It was defended on

March 17th 2015

and approved by

Dr. James B. D. Joshi, School of Information Sciences, University of Pittsburgh

Dr. David Tipper, Graduate Telecommunications and Networking Program

Dr. Prashant Krishnamurthy, Graduate Telecommunications and Networking Program

Dr. Konstantinos Pelechrinis, Graduate Telecommunications and Networking Program

Dr. Adam J. Lee, Department of Computer Science, University of Pittsburgh

Dissertation Director: Dr. James B. D. Joshi, School of Information Sciences, University of Pittsburgh

LOCATION-BASED SOCIAL NETWORKS: LATENT TOPICS MINING AND HYBRID TRUST-BASED RECOMMENDATION

Xuelian Long, PhD

University of Pittsburgh, 2015

The rapid advances of the 4^{th} generation mobile networks, social media and the ubiquity of the advanced mobile devices in which GPS modules are embedded have enabled the location-based services, especially the Location-Based Social Networks (LBSNs) such as Foursquare and Facebook Places. LBSNs have been attracting more and more users by providing services that integrate social activities with geographic information. In LBSNs, a user can explore places of interests around his current location, check in at these venues and also selectively share his check-ins with the public or his friends. LBSNs have accumulated large amounts of information related to personal or social activities along with their associated location information. Analyzing and mining LBSN information are important to understand human preferences related to locations and their mobility patterns. Therefore, in this dissertation, we aim to understand the human mobility behavior and patterns based on huge amounts of information available on LBSNs and provide a hybrid trust-based POI recommendation for LBSN users.

In this dissertation, we first carry out a comprehensive and quantitative analysis about venue popularity based on a cumulative dataset collected from greater Pittsburgh area in Foursquare. It provides a general understanding of the online population's preferences on locations. Then, we employ a probabilistic graphical model to mine the check-in dataset to discover the local geographic topics that capture the potential and intrinsic relations among the locations in accordance with users' check-in histories. We also investigate the local geographic topics with different temporal aspects. Moreover, we explore the geographic topics

based on travelers' check-ins. The proposed approach for mining the latent geographic topics successfully addresses the challenges of understanding location preferences of groups of users. Lastly, we focus on individual user's preferences of locations and propose a hybrid trust-based POI recommendation algorithm in this dissertation. The proposed approach integrates the trust based on both users' social relationship and users' check-in behavior to provide POI recommendations. We implement the proposed hybrid trust-based recommendation algorithm and evaluate it based on the Foursquare dataset and the experimental results show good performances of our proposed algorithm.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

Foremost, I would like to express my gratitude to my PhD advisor, Dr. James Joshi, for his support and encouragement to commit to research during the past six and a half years. I greatly appreciate all his contribution of time and guidance to make my PhD experiences productive and simulating. I would like to thank my dissertation committee members, Dr. David Tipper, Dr. Prashant Krishnamurthy, Dr. Konstantinos Pelechrinis and Dr. Adam J. Lee for their time and suggestions on my research proposal and dissertation writing. I am also grateful to the faculty members that provided great advice towards my future career including Dr. James Joshi, Dr. David Tipper and Dr. Yi Qian.

The members of LERSAIS Lab have contributed immensely to my professional time in Pittsburgh. I would like to thank Lei Jin, for his helpful discussions on this dissertation and the collaboration in other Location-based Social Network related projects. I also want to acknowledge other current and previous LERSAIS members, Amirreza Masoumzadeh, Hassan Takabi, Jesus Gonzales, Nathalie Baracaldo, Saman Taghavi-Zargar and Yue Zhang, for sharing a productive and friendly working environment during the past several years. I cherish the friendship I have had here and wish them all the best. I am also grateful to the faculties and staffs of the iSchool for their support on numerous occasions.

I also acknowledge that the research presented in this dissertation has been supported by the financial assistance provided through the iSchool at University of Pittsburgh, and US National Science Foundation awards IIS-0545912 and DUE-0621274.

I especially thank my beloved Mother and Grandfather for their endless love and support while I am pursuing true knowledge and career path overseas. Last but not least, I am extremely grateful to my husband, Xiben Li, who has been a source of endless love over the past years. His constant support has provided me with the greatest courage to overcome

every challenge in my research and in my life. I have enormous gratitude and heartfelt thanks to all my family for their constant supportive, and this dissertation is dedicated to them.

1.0 INTRODUCTION

Recently, the rapid advances of high-speed wireless mobile networks, the ubiquity of the advanced mobile devices (*e.g.*, smart phones) in which GPS modules are embedded and the powerful interfaces supporting map services such as *Google Maps*, *Microsoft Bing Maps* and *Yahoo! Maps* have greatly enabled and promoted various location-based services. As a combination of the location-based services and social media, Location-Based Social Networks (LBSNs, also called Geographic Social Networks or GeoSocial Networks), such as Facebook Places [2] and Foursquare [3] have been growing rapidly and attracting a huge number of users. For instance, by April 2012, there were 200 million Facebook users posting information including location information [4]. Foursquare community has more than 55 million people worldwide with over 6 billion check-ins [5]. People use such LBSNs as they integrate the functionalities of both social networking and location-based services. LBSN users can explore interesting venues such as restaurants, museums, popular bars, department stores with specials (*e.g.*, discounts, coupons), *etc.*, around their current locations. In a LBSN, users can also add other users as their friends just as they do in traditional social networks and they can also create friendship links with users in the nearby venues.

One of the most popular LBSNs is Foursquare. Foursquare users can create and add a venue/Point of Interest (POI; we use both terminologies in this dissertation) in Foursquare and they can click "Check in" at the venues where they currently are by using a Foursquare mobile App and share such information with others. Users can choose either public "Check in" or private "Check in", and they can earn points by checking in. A user can explore venues around him to find interesting ones and view other people who are checked in there. Users can also easily add friends as they usually do in traditional social networks such as Facebook. They can also choose to post their Foursquare check-ins on Facebook or Twitter

in order to share their location information or activity information with more users. They can choose to be notified of their friends' check-ins. Users can get coupons, discounts and badges [6] by checking in, and this strategy greatly motivates users to check in [7]. Thus, Foursquare accumulates large amounts of information related to the users' personal activities with associated location information.

Exploring and analyzing the Foursquare dataset is very helpful for Foursquare users–both business owners and customers. It is expected that 59% users will search Foursquare for new local businesses [8], thus a business owner will likely miss good opportunities to reach out to customers if his business is not integrated with Foursquare. That is, he may not view his venue statistics of the check-ins on the basis of a certain time period (*e.g.*, one day, last week or last 30 days, *etc.*). He may also lose valuable data such as social reach (*i.e.*, how many users post their check-ins to their Twitter or Facebook account) and the gender and age information of the customers, *etc.* [9]. However, such information is very important and valuable for a business owner to design special offers or prepare for the rush hours. Creating a venue in Foursquare and claiming the ownership of the business will enable the owner to directly obtain the statistics of his own business but the owner still cannot get the statistics of the other venues and user check-ins around his venue. Besides, it is still unknown if there are some features could promote the venue popularity. For LBSN users who would like to explore interesting venues, overall venue popularity also implies good reputation, so it is a very important and valuable feature for customers to choose venues. Thus, the quantitative analysis of the Foursquare dataset will help to understand the general preferences of the Foursquare users. Moreover, sometimes we are more interested in the popular venues among a certain group of users, *e.g.*, a university freshman may be very interested in the cafes that are particularly popular among the freshmen in the same University. Thus, investigating the venue popularity among certain group of users is also an interesting topic. Last but not the least, studying and exploring the interesting venues for individual users are also crucial.

In this dissertation, we first study the overall venue popularity based on the information from the online population, *i.e.*, we employ empirical approaches to analyze the popular venues, popular categories, the temporal and spatial features of the popular venues. Then, we use the probabilistic graphical model to mine the dataset to investigate the latent geographic

topics of users' check-ins; the venues in the same latent topic reflect the preferences of a certain group of users. Lastly, we propose a hybrid trust-based venue recommendation approach for an individual user to find venues that may be interesting for him.

1.1 CHALLENGES

In this dissertation, we aim to validate the following hypothesis:

- Users' check-in activities and venue information in LBSNs can be used to understand users' general preferences about locations and can be used to evaluate the venue popularity.

- Users' check-in activities indicate the human mobility pattern and by appropriately mining the users' check-ins we can discover the human mobility-driven geographic topics, which indicate preferences of location for different groups of users.

- Users' social relationship and check-in behavior can be used to provide good POI recommendation.

Towards this, we work on understanding the users' preferences of locations from the viewpoints of the entire online population, certain groups of users and individual users, respectively. We address research challenges corresponding to the three different types of demographic information (*i.e.*, entire online population, groups of users and individual users) in this dissertation. Since the geographic distance will limit the user activities and most users' activities are within a certain geographic region, the analysis of a local city's dataset can tell more about user activities than the analysis of a general globalized dataset. Therefore, in this dissertation, we focus on the Foursquare dataset pertaining to the greater Pittsburgh area to explore the challenges as follows.

1.1.1 Challenge I: Understanding local venue popularity in LBSNs from the viewpoint of entire online population

Understanding the local venue popularity is important, as it reflects the entire population's preferences regarding the geographic locations. With the LBSN dataset, we are interested in the questions such as "what are the popular venues?", "what are the popular venue categories?" and "what are the temporal and spatial features of the popular venues and what features make the popular venues popular?". These are very common questions related to understanding the human preferences of locations. Thus, in this dissertation, we first would like to answer such questions through the quantitative analysis based on the real dataset, towards understanding human preferences of locations. Moreover, with regard to the popular venues, we are interested to see if there are similar spatial properties among them, or whether there are some specific features that make them more popular, and so on. Thus, analyzing and mining the LBSN data to find out the spatial properties and the features promoting venue popularity are also important. Besides, answering such questions is helpful for designing business and marketing strategies for business owners.

Therefore, the first challenge towards understanding human preferences of geographic locations is to study the venue popularity based on the entire online population's activities, and we will present our work in addressing this challenge in Chapter 3.

1.1.2 Challenge II: Exploring latent geographic topics based on the movements of unobserved groups of users

Our work corresponding to the aforementioned challenge will provide a general overview of the human preferences of locations. However, a further question is how can we understand location preferences of various groups of users? This is because sometimes the users belonging to the same community may have similar tastes and preferences with regard to locations. However, the overall venue popularity derived from the entire online population's check-ins cannot provide a lot of useful information about it. For examples, the overall venue popularity can not tell us what are the clusters of venues that college students like to visit, or what venues sports fans usually visit after watching a sports game? With only the overall

venue popularity, it is very hard to answer such questions. Thus, towards understanding the human preferences of geographic locations, the next challenge is to understand certain groups of users' preferences of locations. The clusters of venues in accordance with the crowd level implies the groups of users' preferences of location and such clusters of venues should not be formed based on the category information or spatial information, but the users' trajectories/check-ins. Thus, the second challenge we would like to address is to employ the probabilistic graphical model to investigate LBSN users' trajectories/check-ins to mine the latent geographic topics of the venues. Also, we are interested in the latent geographic topics formed at different temporal aspects (*i.e.*, on weekdays and at weekends) and those based on the trajectories/check-ins of some special users (*i.e.*, travelers).

1.1.3 Challenge III: Personalized hybrid trust-based POI recommendation

The empirical analysis of the LBSN dataset helps in understanding the venue popularity and the latent geographic topics present a dynamic approach to cluster venues based on the users' check-in data. Both approaches can benefit LBSN users in finding their interesting venues. Furthermore, as LBSN systems usually have heterogeneous venues but there is no straightforward rating mechanism[1] for such venues in most LBSNs [10], it is not easy for users to identify the right venues with good reputation in LBSNs. Even in some platform with the rating of venues, the ratings are not personalized, thus the personal preference of venues are still unavailable. Venue recommendations in LBSNs, thus, has become a challenging research topic in the literature [11] and is the last challenge we address in this dissertation.

1.2 DISSERTATION OVERVIEW AND CONTRIBUTIONS

In order to address the aforementioned challenges, we propose a quantitative analysis to understand the venue popularity in LBSNs, a probabilistic graphical model to mine the geographic topics based on the user check-in records, and a hybrid trust-based POI recom-

[1] In Foursquare, some venues have a rating, but not all the venues have such features.

mendation approach in LBSNs in this dissertation. In particular, the main contributions corresponding to the proposed approaches are as follows:

- We analyze the overall popularity of venues and the overall popularity of venue categories by examining the cumulative number of check-ins and the cumulative number of users who checked in at the venues. Intuitively, the larger the cumulative number of check-ins at a venue or the cumulative number of users who have checked in at the venue, the more popular the venue would be and our first empirical study is based on it. Since Foursquare defines a hierarchical structure of the venue category, we also investigate such category information of the venues to see what kind of venues are popular in LBSNs. Moreover, we investigate Foursquare features such as specials, web presence and menu to analyze if they help attract more check-ins or more users to check in. We also study the influence of the specials on trending venues. After that, we choose the venues in Food and Nightlife Spot categories as examples to examine if the attributes such as *URL*, *Twitter ID* and *Menu* are helpful in making trending venues. Lastly, we explore if the trending venues remain popular during a certain period of time. The proposed quantitative approach to investigate the venue popularity on LBSN dataset implies the entire online population's preferences of the venues on LBSNs. The results could be used to help business owners to improve their business strategies on LBSN as a social media platform.
- We propose to adopt Latent Dirichlet Allocation (LDA) approach to mine the latent geographic topics in LBSNs. A geographic topic is a set of venues that co-exist in many users' check-in records, which implies a certain kind of intrinsic relationship among these venues. These relationships do not depend on the category information or physical geographic distances, but only depend on users' trajectories. The latent topics are very helpful in discovering the clusters of venues based on users' mobility behavior, and can be used in venue recommendation and check-in prediction. By employing the proposed probabilistic graphical model on the Foursquare dataset related to the greater Pittsburgh area, the user mobility-driven geographic topics are obtained. Since the human mobility pattern is expected to be different on weekdays and at weekends, we also investigate the topics on these two different temporal aspects. We study the differences of the latent geographic topics that are based on weekday data and those based on weekend data.

Beside that, we are also interested in the geographic topics formed by travelers. Thus, we extract travelers from the dataset and we study the latent geographic topics based only on travelers' check-ins. The results of this study would help city design center better understand the mobility patterns of local users and travelers and then design/plan better services for them, respectively.

- We propose and implement a hybrid trust-based POI recommendation for LBSN users in this dissertation. The proposed trust-based model is a hybrid trust model, which is based on both the graph-based trust model and the interaction-based trust model. The graph-based trust model illustrates how the trust is affected by the social network structure. The trust computation between users over the social network is done through the unweighted/weighted edges of the graph. Moreover, the check-in activities in LBSNs will also impact the trust of the LBSN users to POIs. The proposed hybrid trust-based POI recommendation will recommend POIs based on the reputation of the POIs from both social and interactions aspects. It is a personalized approach since for each person, the recommendation will be made based on his social network and his check-in history. We also implement the proposed approach and evaluate it using our Foursquare dataset pertaining to the greater Pittsburgh area. The experimental results show the effectiveness of the proposed approach. The proposed recommendation approach can be integrated into the existing LBSN platforms for recommending POIs to LBSN users. Also, the approach can be used in predicting the users' check-in by considering the temporal features.

1.3 SCOPE OF THE DISSERTATION

As we mentioned earlier the users' mobility pattern and mobility behavior will be limited by the geographic distances, thus the work in this paper is focused on a limited geographic area. Although our work is based on the Foursquare dataset of greater Pittsburgh area, the approaches and methods can be used in other local datasets. However, more work needs to be done and more features should be considered when employing the proposed work in a

global dataset, which is out of scope of this dissertation.

1.4 ORGANIZATION

The remainder of this dissertation is organized as follows. In Chapter 2 we review the closely related work in the literature to this dissertation. In Chapter 3, we present the results of the quantitative analysis about venue popularity of our dataset pertaining to Greater Pittsburgh area. We also explore the relation between Foursquare features and the venue popularity, as well as the temporal and spatial properties of the popular venues. Such analysis and result are very helpful in understanding the preferences of the entire online population. In Chapter 4, we introduce how to employ the Latent Dirichlet Allocation model in our LBSN dataset and present the results of latent geographic topics by exploring the users' check-in data. The latent topics reflect the location preferences of unobserved groups of users. We also discuss the latent topics and how to use these topics in applications. In Chapter 5, we present the hybrid trust-based POI recommendation approach for individual users. We implement the approach, evaluate it by using our dataset and present the results. In Chapter 6, we present contributions, limitations and future directions related to the dissertation .

2.0 BACKGROUND AND RELATED WORK

In this chapter, we first describe the terminology used in the proposal. Then we present closely related work to LBSNs in the literature.

2.1 TERMINOLOGY

The terminology used in this proposal is summarized in this section.

Venue/POI. A Foursquare venue/POI is a physical location. It can be a place of business office or private residence where Foursquare users can check in.

Venue Category. Foursquare defines a hierarchical list of categories applied to venues. There are 9 top categories in the hierarchical structure and they are: *Arts & Entertainment*, *College & University*, *Food*, *Professional & Other Places*, *Nightlife Spot*, *Great Outdoors*, *Shop & Service*, *Travel & Transport* and *Residence*. There are some venues in Foursquare which do not have category information and we do not consider these venues in this dissertation.

Specials. There are eight different types of specials provided by Foursquare and they are: *mayor*, *count*, *frequency*, *regular*, *friends*, *swarm*, *flash* and *other*. They could be used for specific promotions to get new customers or give rewards to the most loyal customers [12]. Third parties, *e.g.* American Express which provides specials like "get $5 off using your American Express card", also can be providers of Specials in Foursquare, but we do not consider them in the dissertation. The provider of the specials in this dissertation is Foursquare.

Trending venue. The trending venues in Foursquare are defined as the venues near the

user's current location with the most people currently checked in [3]. They are usually the popular venues at a certain time. The trending venues are provided by Foursquare API and during our data crawling period we found that a venue in Greater Pittsburgh area was considered a trending venue if there were at least 5 users currently checked in.

Menu. It is a Foursquare feature that provides the menus in venues in Food and Nightlife Spot categories to users.

Web Presence (*URL* and *Twitter ID*). In Foursquare, a venue's owner can provide a URL of his website and a Twitter ID that links the owner's tweets in Twitter. Thus, Foursquare users can view the website to get more information about the venue. They can also tweet messages about the venue or view the tweets about the venue.

2.2 RELATED WORK

2.2.1 Analysis related to Human Mobility

Understanding the human mobility has recently become a hot research area, *e.g.*, exploring the scales of human mobility [13], mobility and migration patterns [14], the geographic constraints on social groups [15] and human mobility's impact on social relationships [16], *etc.*. A lot of work employ mobile phone datasets to investigate the human mobility [17, 18, 19]. González *et al.* study the six-month mobile phone users' trajectories in [20]. Their results show that human trajectories have a high degree of temporal and spatial regularity. That is, human mobility follows simple reproducible patterns, in spite of the diversity of their travel history.

Noulas *et al.* analyze the user activity patterns using Foursquare dataset in [21]. Then, in [22], they study the urban mobility patterns of people in several metropolitan cities. Although the human movements vary in different cities, they identify a general pattern for human mobility.

Li and Chen's work is the first large-scale quantitative analysis of a real-word LBSN service [23]. They investigate user profiles, update activities, mobility characteristics, social

graphs, and attribute correlations. Their work is very general in these aspects and our proposed work is different from theirs in that we focus on venues and users' preferences of the venues.

In [24], Allamanis *et al.* study the temporal evolution of LBSNs. They investigate the Gowalla dataset to explore how the social links are created, how the social triangles are created and how users' mobility patterns affect their social influence. Based on the findings, they also define network growth models to reproduce the spatial and social properties which can describe the fundamental mechanisms affecting user behavior.

Noulas *et al.*'s work in [25] uses local dataset (*i.e.*, New York) and analyzes the venues according to different categories. However, this work aims to explore the semantic annotations for clustering geographic areas and it still uses the data collected from Twitter.

Scellato and Mascolo's work measures the user activity in a LBSN [26]. Their work studies how Gowalla users connect with their friends and how users check in at different places. Their work shows that the double Pareto-like distribution can describe the distribution of the number of friends, and the log-normal distribution can describe the numbers of check-ins and places. Their work is mainly based on the number of friends and the number of check-ins of a venue, thus they do not study the venues in detail and they do not examine the popular venues.

2.2.2 Data Mining in LBSNs

Cho *et al.* study the relationship between mobility and friendship by exploring the user movements in LBSNs in [27]. Their work show close connection of them, *e.g.*, 10% to 30% of all human movements can be explained by social relationship and 50% to 70% human movements are related to periodic behavior.

Cheng *et al.* explore the check-ins to analyze human mobility patterns in spatial, temporal, social, and textual context [28]. First, their work uses the global data collected from Twitter. They do not present the cumulative information and the category information of the venues as in this dissertation. Thus, their analysis about check-ins in spatial-temporal aspect is not as detailed as ours. Moreover, not every Foursquare user has a Twitter account.

In our dataset, only 28.44% of the users have Twitter IDs. Thus, many Foursquare users are excluded in their dataset. Their work focus on the human mobility patterns and our work is different from theirs as we focus on the users' preferences of venues.

Gao *et al.* explore social-historical ties in LBSNs by studying the user check-in behavior in [29]. They use both power-law distribution and short-term effect to capture users' historical check-ins. Their experimental results demonstrate how social and historical ties can help location prediction.

Vasconcelos *et al.*'s work investigates the tips, dones and to-dos information in Foursquare [30]. This work uses a global dataset and it focuses on the user interactions by posting tips and by marking them as done or to-do.

Ferrari *et al.* employ LDA to extract the urban patterns from location-based social networks in [31]. Our prposed work is different from theirs mainly in two ways. First, our topics are based on the users' check-ins, *i.e.*, we use a user's check-ins as a document and a venue in a check-in as a word in the LDA model. However, their work uses a venue and a time slot as a word and forms a document of all check-ins in a day in their LDA model. Thus, their work focuses on the human mobility patterns at different times within a city and our proposed work is more human centric than theirs as we investigate the topics based on trajectories of large groups of users. The data set used in their work is crawled from Twitter but not from Foursquare directly. In Foursquare, a user can use his Twitter account to login and post his check-ins on Twitter, but not every user has a Twitter account and neither is every user likely to post his check-ins on Twitter.

Ferrari and Mamei also investigate the topics based on the user's trajectory in [32]. The data set in [32] is the daily whereabouts of two persons over the period of almost one year. They divide a day into 48 time slots and each time slot lasts for 30 minutes, so the 48 places each day form a document. Thus, their work is still time based topics, which focuses on a single user's mobility pattern at different times and is thus very different than our approach.

Farrahi and Gatica-Perez's work in [33] also use LDA to discover the routine behaviors. They use the Reality Mining data set [34, 35] that contains a one-year mobile phone sensor data recording 97 subjects from 2004 to 2005. The routine behaviors in their work are still time based topics, which are different from ours, as we do not consider the temporal factors

in our model. Besides, the locations in their work are simply labeled by "Home", "Work", "Other" and "No Reception". Thus, the rich POI information is lost in their work.

Yuan *et al.* also propose a framework to discover regions of different functions in a city by using human mobility with regard to both regions and POIs in the region in [36]. Their topic model is based on LDA and Dirichlet Multinomial Regression. In their work, they use the GPS trajectory datasets. Besides, their work aims to discover the regional topics, which is different from ours.

Cranshaw and Yano employ LDA to distill the proto-neighborhoods from Foursquare data set in [37]. In their work, the word is the category of the venue and the document is the check-ins in a region. Regions are small grids that divide space according to the latitude and longitude space. Thus, each region can be described by the topics, which can help to understand the neighborhood.

Chang and Sun analyze users' check-ins in general in Facebook Places in [38] and LDA is also used in their work to investigate the user membership in a low-dimensional representation of the places. Since their work is in general about the check-in analysis so the topic model is only a small part and they only give three topics without analyzing the topics in details as we do. Besides, they do not consider the differences in mobility patterns of users in the weekdays and weekends.

In [39], Zheng *et al.* mine interesting locations and the classical trajectories based on travelers' GPS trajectories. Such information can be used to understand the correlation between users and POIs, thus enabling POI recommendations to travelers. They first propose to use the Hyperlink-Induced Topic Search (HITS) model to extract the location interests [40]. However, their approach does not consider the social influences among the users, as they use the GPS trajectories of individual users and it is very hard to track the social connections among these users. Based on this work, Zheng *et al.* propose collaborative POI and activity recommendation approaches with GPS data in [41]. By using the POI features and activity-activity correlations in their proposed approach, they demonstrate improvements on both POI and activity recommendations over the simple baseline. Nevertheless, they still do not explore the social influences in their approach.

Joseph *et al.* also employ LDA to mine the LBSNs dataset to cluster users with latent

topics in [42]. Their work employs the same mapping approach and LDA as our work in Chapter 4 but we mine the different dataset compared with theirs[1]. Their work mine the Foursquare dataset collected through Twitter and our dataset is collected from Foursquare directly. Both of our work obtain the similar interesting results: they find interest factor driven cluster–Sport Enthusiast cluster and we also find sports related topic; they find community driven clusters–Gay-bar cluster and we find the UPMC related topic; they find user type factor driven cluster–a cluster seems to consist of Stanford students and we also fine the clusters that consist of PITT and CMU students, respectively. Moreover, our work investigates the latent geographic topics based on different temporal aspects (*i.e.*, weekdays and weekends), which characterizes the differences of human mobility patterns between weekdays and weekends.

2.2.3 Trust in Online Social Networks and Location Recommendation in LBSNs

Sherchan *et al.* review trust models in social networks in [44]. In the survey, they present the properties of trust and social capital. They roughly categorize the social trust computation approaches as network-based trust models, interaction-based trust models and hybrid trust models.

The network structure/graph-based trust models usually leverage the social network structure to the level of trust in social networks [45, 46, 47, 48, 49]. In contrast to the network structure/graph-based trust models, the interaction-based models only use interactions within the network for trust computation [50, 51, 52, 53]. However, both social network structure and interactions are very important and both of them should be considered in social trust computation. Trifunovic *et al.* propose such a hybrid trust model for opportunistic networks in [54]. The literature on hybrid social trust models is limited and in this dissertation our proposed trust-based recommendation approach is a hybrid trust model.

Ye *et al.*'s work does an initial analysis about Foursquare in [55]. But the dataset in this work is global, not the local one as in our proposal. They propose a friend-based

[1]Both of their work and our publication [43] about using the LDA to mine the human mobility-driven geographic topics are published in LBSN 2012.

collaborative filtering (FCF) approach and Geo-Measured FCF for location recommendation. Another work from them also study the physical distance's influence on the users' check-ins. But their main work is also the POI recommendation on global dataset [56] .

Berjani and Strufe propose a personalized regularized matrix factorization based recommendation for POIs in LBSNs [10]. They argue that the main challenge of POI recommendation in LBSNs is the lack of straightforward rating of POIs and propose a user preference approach based on check-in counts to address this issue. They evaluate their proposed scheme on Gowalla dataset and the results confirm the feasibility of making POI recommendations by using CF-based techniques.

Leung *et al.* propose a collaborative location recommendation (CLR) framework based on GPS data in [57]. They classify users into three categories: Pattern users, Normal users and Travelers by the entropies of the locations that they have visited. They also employ the Community-based Agglomerative-Divisive Clustering (CADC) algorithm to cluster users, locations and activities. Based on the clusters, the CLR can generate more precise and refined recommendations.

Zheng *et al.* propose to recommend friends and locations based on the location history in [58].Their recommendation is based on calculating the similarity of the users' location histories and their work is based on the GPS dataset, which does not contain the social graph. In [59], Bao *et al.* study the user preferences and social opinion in location recommendation. They use the tips information on Foursquare to model the user preferences in their work. However, their work still does not consider the social relationships among the users.

In [60], Ying *et al.* propose an Urban POI-Mine (UPOI-Mine) approach to recommend POIs by considering users' preferences and location features simultaneously. They consider the social factors, personal preferences based on the category context, highlight context and the popularity of the POIs in making recommendations.

Scellato *et al.* study the socio-spatial features in LBSNs and they investigate the place features in link prediction in LBSNs in [61, 62]. Noulas *et al.*. extract and study the LBSNs users' mobility features for predicting next check-ins in [63]. They propose several features to capture the factors which may drive users' future check-ins. Then, they propose a random-walk based new POI recommendations in LBSNs in [11]. They explore the frequency of visits

to new POIs by LBSN users and study the assumptions of using web-filtering algorithms in human mobility predictions in the proposal. Based on the analysis of the LBSN dataset, they show that state-of-the-art filtering algorithms do not produce high quality recommendations and propose personalized random walk based recommendations. Their experiments show a 5%-18% improvement using the proposed random-walk recommendation on Gowalla dataset.

3.0 VENUE POPULARITY ANALYSIS

In this chapter, we study the venue popularity and the features that could impact the venue popularity in Foursquare. Our quantitative analysis results could be used to help business owners to improve their business strategies. We use an empirical approach to analyze the popular venues of our dataset and we also investigate category information of the venues to see what kind of venues are popular in LBSNs, since Foursquare defines a hierarchical structure of the venue category. Moreover, we study the local hot spots that indicate people's preferences of geographic locations and we explore if the Foursquare features such as special offers and web presence would help venues become more popular in general. Then we focus on trending venues (*i.e.*, the popular venues at a certain period) to investigate the influence of these features on venue popularity at different times. In this chapter, we present a comprehensive analysis of the venue popularity, in order to understand the entire online population's preferences of the venues.

The rest of this chapter is organized as follows. In Section 3.1, we introduce our data collection strategy and the summary of the dataset used in this chapter. In Section 3.2, we present our analysis of the overall venue popularity in details. In Section 3.3, we focus on trending venues and investigate their spatial and temporal properties. We also study the Foursquare features on promoting the temporal venue popularity and the hotness status of the trending venues. In Section 3.4 we discuss some interesting findings and in Section 3.5 we summarize this chapter.

3.1 DATA COLLECTION STRATEGY & SUMMARY

The data was collected in a square region that centered at Pittsburgh downtown and with length of a side of around 40 miles. We first used Foursquare APIs [64] to discover as many venues as possible in this region. In total we got 70,390 venues belonging to 9 top categories and 271 second-level categories as defined by Foursquare. In this dissertation we collect several different datasets for serving our goals by monitoring check-ins at these venues. We briefly summarize our data collection strategy here and we will overview the datasets in each corresponding chapters in detail.

- *Venue Status Dataset and Trending Venue dataset*: we collect the venue status information daily and collect trending venues hourly. They are used for venue popularity analysis in Chapter 3.
- *User Check-in Dataset*: we collect user check-ins by monitoring check-ins at venues hourly and this dataset is used in mining latent topics in Chapter 4.
- *User Check-in and User Friendship Dataset*: Besides the user check-in information, we also collect the user friendship information in LBSNs and we use both of them in Chapter 5.

In this chapter, we collected the following venue status information in the period between Feb. 23 and Apr. 24, 2012:

- *Cumulative number of Check-ins* in a venue: It represents the total number of check-ins ever made at this venue till a specific date (*i.e., data collection date*).
- *Cumulative number of Users* in a venue: It represents the total number of users who have ever checked in at this venue till a specific date (*i.e., data collection date*).
- *Specials* of a venue: When the venue has specials within a period, we collect the information of these specials. Not every venue has such features.
- *Menu*, *URL*, *Twitter ID* of a venue: When the venue has any of these resources, we collect the corresponding information. Not every venue has such features.

We are also interested in the trending venues in greater Pittsburgh area. In order to collect such information we set Pittsburgh downtown as the center and define a circle with

radius of 25 miles[1]. We then gathered the trending venues per hour in this circle. For a trending venue at a certain time, we obtained the number of users who were currently at the venue.

The summary of our dataset corresponding to the two strategies used in this chapter is shown in Table 3.1. In essence, there are 66,993 venues with at least one check-in since they were created and the other 3,397 venues have no users checked in after they were created. We also get 10,087 records for the trending venues and these records are from only 603 distinct venues. Please note that, in this dissertation, when we mention venues in Foursquare we mean venues in Pittsburgh area in Foursquare, unless mentioned otherwise.

Table 3.1: Summary of our Foursquare dataset used in Chapter 3

Number of venues	70,390
Number of venues with at least one check-in	66,993
Number of trending venue records	10,087
Number of distinct trending venues	603
Number of specials records	6,501
Number of venues with specials	187
Number of venues with URL	2,779
Number of venues with Twitter ID	2,831
Number of venues with menu	1,217

3.2 OVERALL POPULARITY ANALYSIS

We are interested in the venue popularity based on the number of check-ins and the number of visitors of a venue, as well as the category popularity. Towards this, we analyze the overall popularity of different venues and different categories by examining the cumulative number of check-ins and cumulative number of users in this section. Then we make a initial study

[1]We remove the trending venues that is not in the 70,390 venues.

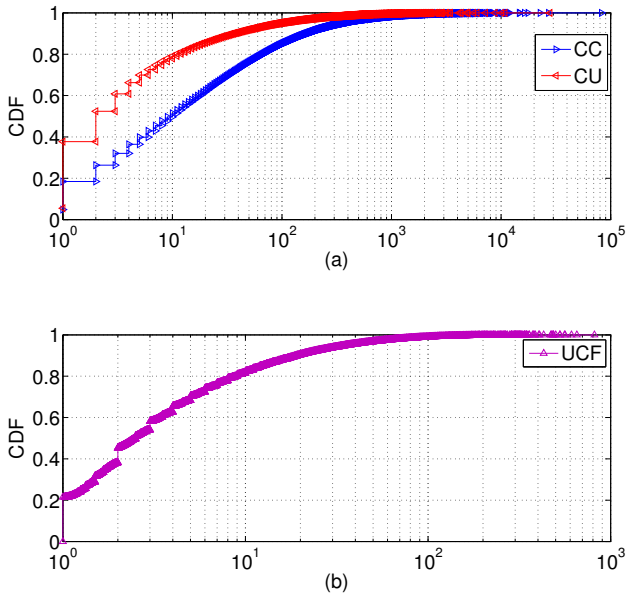

Figure 3.1: (a) CDF of CC and CU; (b) CDF of UCF

about the Foursquare features such as special, web presence and menu to see their impacts on improving the popularity of venues.

3.2.1 Overall Venue Popularity

It is natural to relate the number of check-ins or the number of visitors at a venue with the popularity of a venue. Thus, we use the following measurements to investigate the popularity of a venue:

a) cumulative number of check-ins at venue i (CC_i)

b) cumulative number of unique users of venue i (CU_i)

c) user check-in frequency of a venue i (UCF_i)

We can get CC_i and CU_i from the venue information in our dataset and we calculate UCF_i as:

$$UCF_i = \frac{CC_i}{CU_i} \tag{3.1}$$

We plot the empirical cumulative distribution function (CDF) of these three measures[2] in Figure 3.1 and we also list the basic statistical features of CC, CU and UCF in Table 3.2. Since the maximum cumulative number of check-ins and maximum cumulative number of users at a venue are large, *e.g.*, 82,590 and 27,944, respectively, we use log scale in x axis.

In Figure 3.1, we can see that the CC of about 85% of the venues is no more than 100 and the CU of about 96% of the venues is no more than 100. Moreover, about 18% of the venues have only one check-in and about 37% of the venues have only one user who checked in there. From Table 3.2, we can see that the CC of 50% of the venues is no more than 10 and the CU of 50% of the venues is no more than 2, which indicates that the CC and CU in Foursquare follows the long tail behavior [65]. We are using the cumulative data to analyze the venue popularity as we aim to investigate the overall venue popularity but not limited to the venue popularity during a certain time period. In [21], Noulas *et al.* analyze the check-ins collected during their data collection period and the complementary cumulative distribution function (CCDF) of the number of check-ins also exhibit heavy tail and power-law behavior. We also study the check-ins for venues in Section 5.4 and show the CCDF of check-ins in Figure 5.5; our results also show that a few venues obtain a large number of check-ins while a larger number of venues have only few check-ins. Moreover, we will focus on trending venues to investigate the venue popularity during a certain period time in Section 3.3.

Generally, the larger UCF indicates the more check-ins from a user on average. In Table 3.2, the max UCF is 815, which is related to a resident address and there is only one person checked in this venue 815 times. In Figure 3.1, the UCF of about 83% of the venues is no more than 10, which means the average check-ins per user at these venues is at most 10. 50% venues have an average check-ins per person of less than 3.

The UCF also implies the users' loyalty to a venue, especially for venues in Food, Shop & Service and Nightlife Spots categories. Thus, UCF of such venues can not only help the business owners to understand their customers, but also help customers to compare

[2]We remove the venues with zero users in calculating the UCF.

Table 3.2: Statistics of the CC, CU and UCF

Min of CC	0
Max of CC	82,590
Mean of CC	94.70
Median of CC	10
Standard Deviation of CC	627.56
Min of CU	0
Max of CU	27,944
Mean of CU	26.87
Median of CU	2
Standard Deviation of CU	218.19
Min of UCF	1
Max of UCF	815
Mean of UCF	8.68
Median of UCF	2.5
Standard Deviation of UCF	21.99

Table 3.3: Top 10 popular venues based on cumulative number of check-ins and cumulative number of users

	AC	*AU*
1	Pittsburgh International Airport	Pittsburgh International Airport
2	CONSOL Energy Center	PNC Park
3	PNC Park	CONSOL Energy Center
4	Ross Park Mall	Heinz Field
5	Rivers Casino	Rivers Casino
6	Heinz Field	Ross Park Mall
7	Robinson Mall	Hofbräuhaus Pittsburgh
8	Giant Eagle Market District	Robinson Mall
9	Lowes Theatre	IKEA
10	Cathedral of Learning	Lowes Theatre

similar merchants which offer similar types of services. For example, if we define UCF as the customer loyalty then the average customer loyalty in Food category in Foursquare is about 3 and the customer loyalty of Tom's bakery house is 5. Based on this, Tom can see that there are many returning customers in his venue and he should offer specials to these returning customers as rewards for their loyalty. Meanwhile, if a customer wants to find a good nearby bakery house, he can compare the total number of customers and the user check-in frequency of the venue with that of the corresponding average values. If both of the parameters of Tom's bakery house are higher than that of the similar other venues, then it indicates that Tom's bakery house is better than the average. However, the assumption here is that all check-ins are honest.

Since the most popular venues attract most of the users, we list the name of the top 10 popular venues with the largest CC and CU in Table 3.3. The top 10 popular venues with the largest UCF are all Residence venues, but we do not list them because of the privacy concerns.

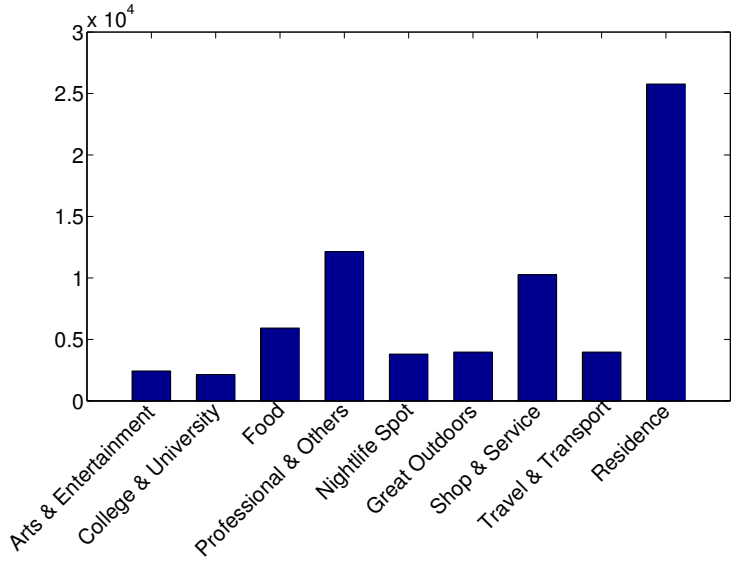

Figure 3.2: Venue distribution in 9 top categories

3.2.2 Overall Category Popularity

In our dataset, there are more than 70 thousand venues, and most of the venues are associated with one or more than one categories (this is because Foursquare defines a hierarchical structure of categories). Thus, we are also interested in what categories people usually check in at. In the Foursquare defined hierarchical category structure, there are 9 top categories and we plot the venue distribution in these 9 top categories in Figure 3.2, in order to present an overview of the venues and categories of our dataset. We can see that Residence has the largest number of venues followed by Professional & Others and Shop & Service.

Based on the measures of the venue popularity, the overall popularity of the category C can be measured by the sum of the cumulative number of check-ins of all the venues in the category (SC_C) and the sum of the cumulative number of users (SU_C) of all the venues in the category:

$$SC_c = \sum_{i}^{N_C} CC_i \tag{3.2}$$

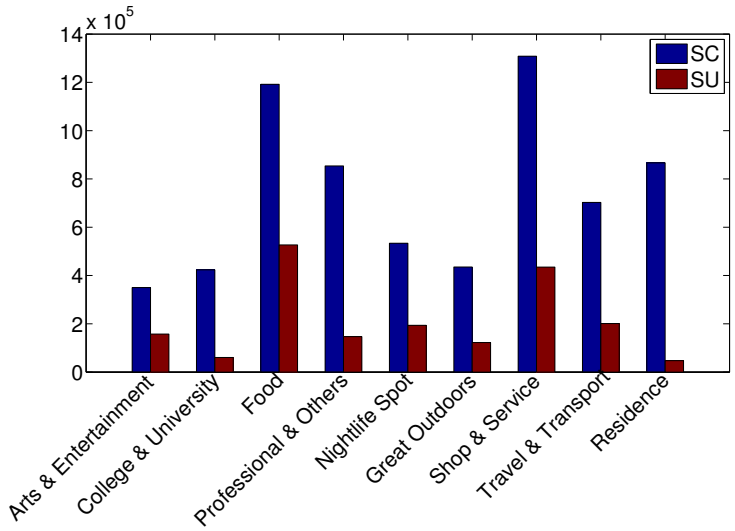

Figure 3.3: Distribution of SC and SU in 9 top categories

$$SU_c = \sum_{i}^{N_C} CU_i \tag{3.3}$$

Here N_C is the total number of the venues in category C. We plot the SC and SU in Figure 3.3. We find that the distribution of SC and SU are different from the venue distribution shown in Figure 3.2. From Figure 3.3, we can see that the Shop & Service category has the largest number of user check-ins, and the Food category has the second largest number of user check-ins. Although the number of venues in Residence category is the largest, the number of check-ins in this category is just the third largest. Also, Table 5.2 based on our user check-in dataset also shows that the most popular categories are Food, Shop & Service, Professional & Other Places and Residence categories.

We also observe that the distribution of the SC_C is different from that of SU_C as shown in Figure 3.3. It can be seen that the Food category has the most number of users checked in followed by that of the Shop & Service category. An interesting observation here is that the Residence category has the least number of users checked in, although it has the largest number of venues. A possible reason is that Residence venues are usually private places, thus people other than the owner of the venues just do not check in at these venues.

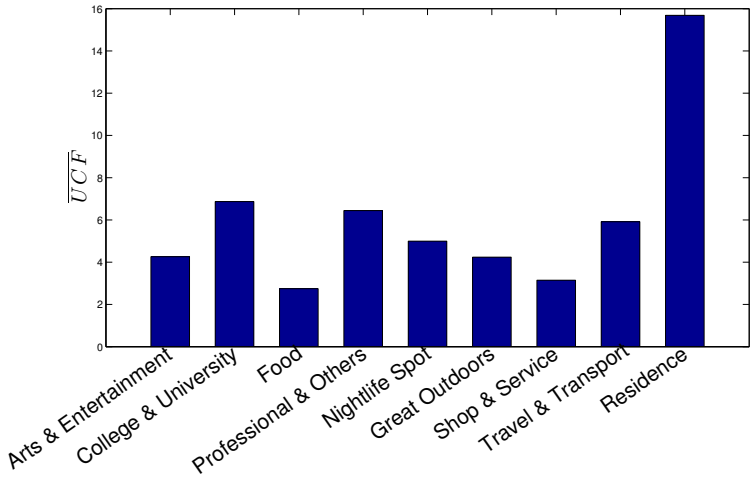

Figure 3.4: Distribution of $\overline{UCF}$ in 9 top categories

We define the average user check-in frequency of category C as:

$$\overline{UCF_C} = \frac{\sum_i^{N_C} UCF_i}{N_C} \tag{3.4}$$

The distribution of $\overline{UCF_C}$ is shown in Figure 3.4. The Residence category has the largest average user check-in frequency followed by that of the College & University and Professional & Others category. Zang *et. al*'s work also shows that home and work are the top 2 locations of mobile users [66]. Since we use the cumulative data, our results imply the higher probability of check-ins in the venues that users usually stay at.

3.2.3 Relation between Foursquare Features and Overall Venue Popularity

In Section 3.2.1 and Section 3.2.2 we quantitatively analyze the venue popularity and category popularity. We are also interested in studying if there is a strong relation between Foursquare features' (*e.g.*, specials, web presence and menu) and the venue popularity. Thus, in this subsection we discuss three such features and investigate if they could help promote overall venue popularity. The following four measures are used in this subsection:

- Check-ins during a time period (CP) with the first day m and the last day n: $CP = CC_n - CC_{m-1}$
- Number of New Users during a time period (NUP) with the first day m and the last day n: $NUP = CU_n - CU_{m-1}$
- Average Daily Check-ins during the time period ($\overline{DC}$) with the first day m and the last day n: $\overline{DC} = \frac{CP}{n-m+1}$
- Average Daily New Users during the time period ($\overline{DNU}$) with the first day m and the last day n: $\overline{DNU} = \frac{NUP}{n-m+1}$

3.2.3.1 Specials During our data collection period, we obtain 6,501 specials records in total. However, there are only 187 distinct venues in these records, thus the average special period per venue is about 35 days ($\frac{6501}{187} \approx 35$). Specials are not always continuous. The venue owner can post specials for just one week or he can post such specials for one month, and they can choose when to post the specials. We use the average daily check-ins ($\overline{DC}$) and the average daily new users ($\overline{DNU}$) and compare them to see if there is any difference between the venue's promotion periods and the periods without promotions. If the $\overline{DC}$ and $\overline{DNU}$ during the promotion periods are more than that in periods without promotions, it indicates that the specials may help the venue in attracting more customers.

We show our methodology as follows. We first pick all the venues which have specials from Mar. 19 to Mar. 25, 2012 in our dataset[3]. Then, we choose the venues which have no specials lasting for one week from the previously picked venues and finally we get 23 venues in total, as we would like to compare these venues in the same time period. Figure 3.5 shows the $\overline{DC}$ of the 23 venues and Figure 3.6 shows the $\overline{DNU}$ of 23 venues. In Figure 3.5, we see that some venues have significant increase in average daily check-ins with specials. The max difference between average daily check-ins with specials and that of one without specials is 4.16 at venue seven. In some other venues, the specials do not seem to attract more check-ins. In Figure 3.6, we can also see that some venues have increase in average daily new users with specials, however the increase is not as obvious as that of the average daily check-ins. In some other venues, the specials do not seem to attract more new users.

[3]We randomly pick this period as an example.

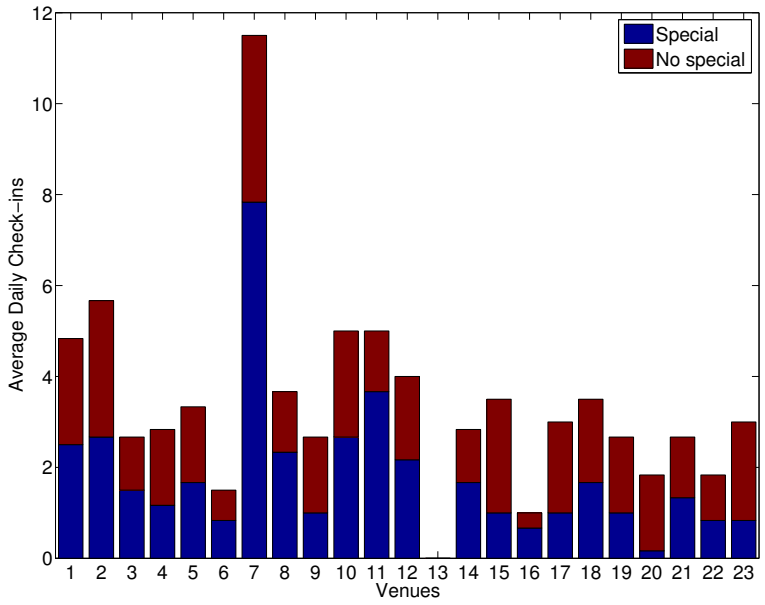

Figure 3.5: Average daily check-ins

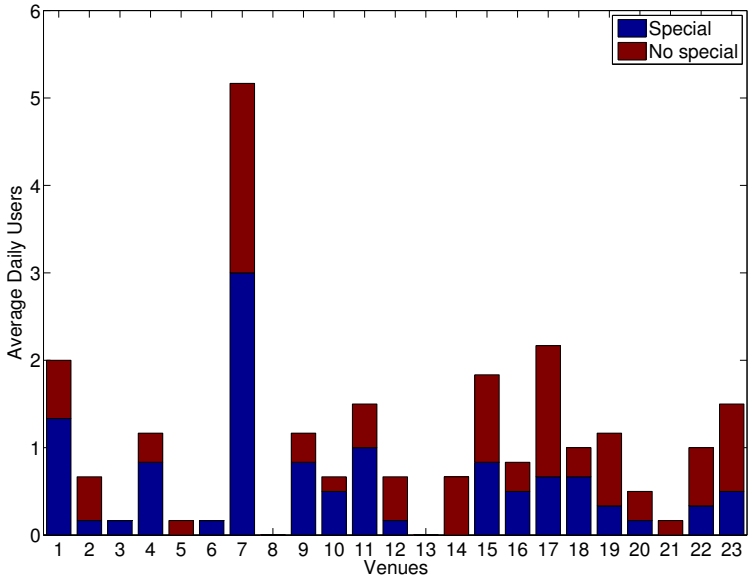

Figure 3.6: Average daily users

Table 3.4: Summary of the CP and NUP related to web presence

	URL	Twitter ID	both	All
Venues	2,779	2,831	1,591	70,390
Sum of the CP	205,557	207,312	142,206	1,432,059
Sum of the NUP	69,284	64,487	45,634	383,641

We can also see that some venues gain more in terms of both the check-ins and new customers when they offer specials. For example, *Bocktown Beer and Grill* (venue seven) gets both the maximum check-ins and maximum number of new customers. In *Bocktown Beer and Grill*, the type of the specials provided is *frequency* and the special message is "*Have you gotten 'FRIED' lately? Add an extra topping to an order of our fresh-cut fries today on us!*". Since users do not check in frequently, specials may be one positive factor for improving the popularity of a venue in Foursquare.

3.2.3.2 Web Presence In this part, we investigate the CP and the NUP related with the venue which has an URL or Twitter ID or both during our data collection period, which is summarized in Table 3.4. We plot the CDF of the CP and the CDF of the NUP of the venues with URL, Twitter ID, both URL and Twitter ID and all venues in Figure 3.7 and Figure 3.8.

From Figure 3.7, we can see that the CP of over 52% of the venues with Twitter ID is more than 10 and the CP of over 60% of the venues with URL or both features is more than 10 in our data collection period. However, the CP of only about 19% of the total venues is more than 10. However, there is no such significant difference for the venues whose CP is more than 1000. Figure 3.8 gives similar results about the NUP. NUP of over 32% of the venues with Twitter ID is more than 10; NUP of over 40% of the venues with URL is more than 10; NUP of more than 42% of the venues with both features is more than 10 in our data collection period. However, NUP of only about 7% of the total venues is more than 10. Thus, there seems a correlation between the venues with good web presence and the

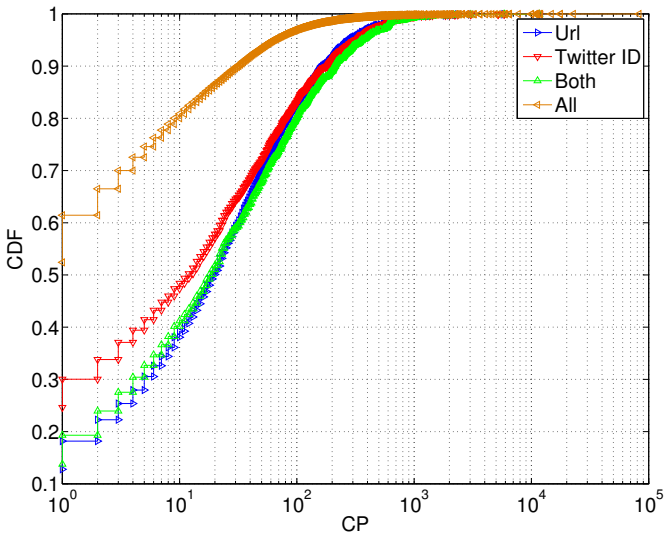

Figure 3.7: CDF of the CP of the venues with URL, Twitter ID, both URL and Twitter ID and all venues

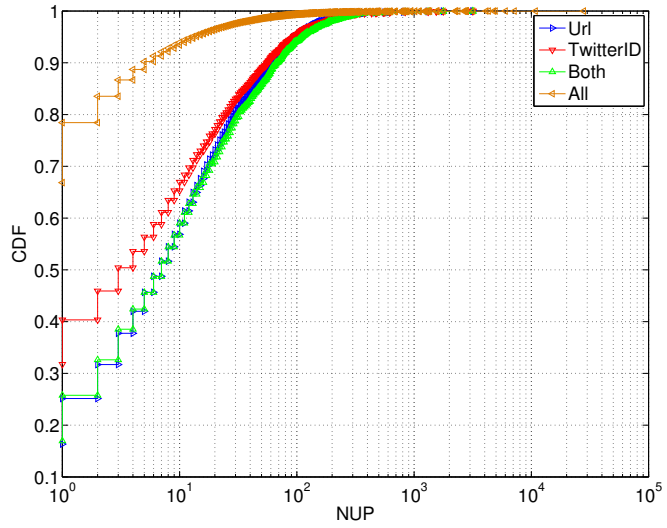

Figure 3.8: CDF of NUP of the venues with URL, Twitter ID, both URL and Twitter ID and all venues

Table 3.5: Summary of the CP and NUP related to menu

	Food	*Nightlife Spot*
Venues with menu	1,150	67
All venues	5,921	3,802
Sum of the CP in venues with menu	73,739	29,841
Sum of the NUP in venues with menu	11,636	5,087
Sum of the CP in all venues	101,793	33,164
Sum of the NUP in all venues	212,759	88,237

venue attracting more users. Moreover, There is still no such significant difference between the cases with features and without features for a venue whose NUP is more than 1000.

3.2.3.3 Menu We explore the CP and NUP of venues which have menus in Food and Nightlife Spot categories. Table 3.5 summarizes the information related to the web menu of the venue. We also plot the CDF of the CP and the CDF of the NUP of the venues with menu and all the venues in Food category and Nightlife Spot category in Figure 3.9 and Figure 3.10.

In Figure 3.9, the CP of about more than 67% of the venues with menu in Food category is more than 10 and the CP of more than 83% of the venues with menu in Nightlife Spot category is more than 10 within our data collection period. However, the CP of only about 37% of the total venues in Food category is more than 10 and the CP of only about 22% of the total venues in Nightlife Spot category is more than 10. However, there is no such significant difference for the venues which have more than 1000 check-ins. In Figure 3.10, the NUP of about more than 52% of the venues with menu in Food category is more than 10 and the NUP of more than 75% of the venues with menu in Nightlife Spot category is more than 10 within our data collection period. However, the NUP of only about 24% of the total venues in Food category is more than 10 and the NUP of only about 9% of the total venues in Nightlife Spot category is more than 10. Thus, there seems a correlation

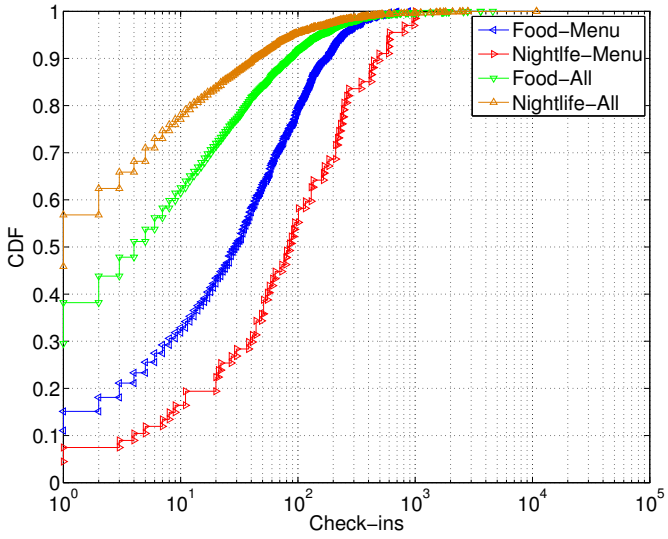

Figure 3.9: CDF of the CP of the venues with menu and all the venues in Food and Nightlife Spot categories

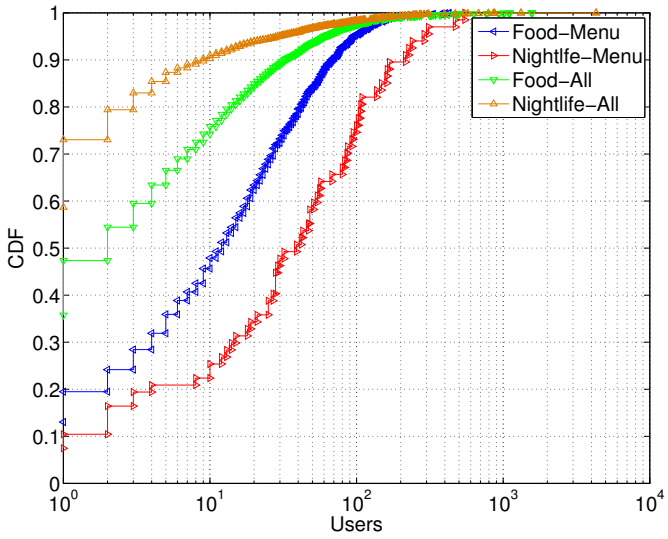

Figure 3.10: CDF of NUP of the venues with menu and all venues in Food and Nightlife Spot categories

between the venues with menu and the venue attracting more users. Moreover, there is still no such significant difference for the venues which have more than 600 new users.

3.3 POPULARITY AT DIFFERENT TIMES

In Section 3.2 we investigate the overall venue popularity, besides that we are also very interested in the venue popularity during a certain time. In this section, we focus on trending venues, *e.g.*, popular venues during a certain time. We first investigate the trending venues based on the spatial and temporal aspects, and then study the influence of the specials on venue popularity over time. After that, we choose the venues in Food and Nightlife Spot as examples to examine if the attributes such as the URL, Twitter ID and menu are helpful in making trending venues. Finally, we explore if the trending venues remain popular over a certain period of time.

3.3.1 Spatial and Temporal Analysis of Trending Venues

We show the physical geographic locations of the trending venues in our dataset in Figure 3.11. It is obvious that most of the trending venues are around Pittsburgh Downtown area. There are also many trending venues along the main highways or the intersections of the freeways.

We plot the distribution of the number of trending venues for each day on a weekly basis in Figure 3.12. The number of trending venues on Saturdays is always the largest. This is possibly because a lot of people may hang out on Saturdays. We can see that the largest number of trending venues occurred on Mar. 17, which was St. Partrick's Day, indicating that more people would likely hang out during holidays. The number of trending venues on Mondays is almost the fewest. There are no significant differences among the numbers of the trending venues on other days. We also summarize the trending venues every hour in Figure 3.13. We can see two peaks in Figure 3.13. One is at 2pm and the other is at 9pm. We can also see that the top four hours with the largest number of trending venues are 9pm,

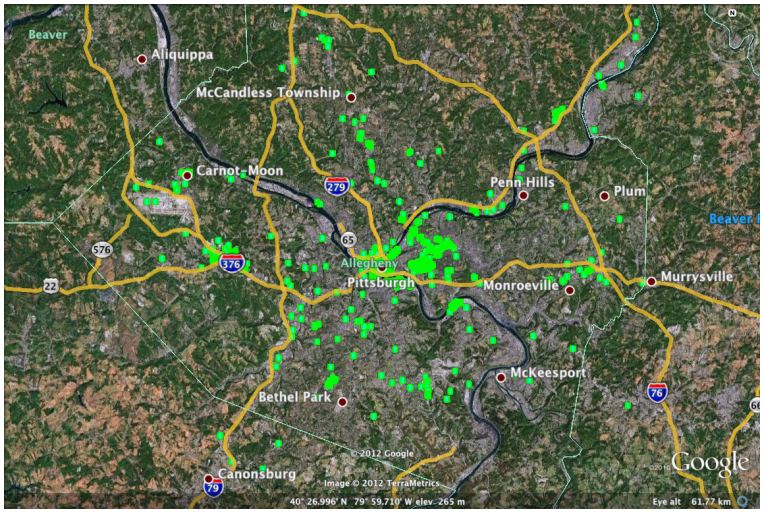

Figure 3.11: Trending venues in Pittsburgh area

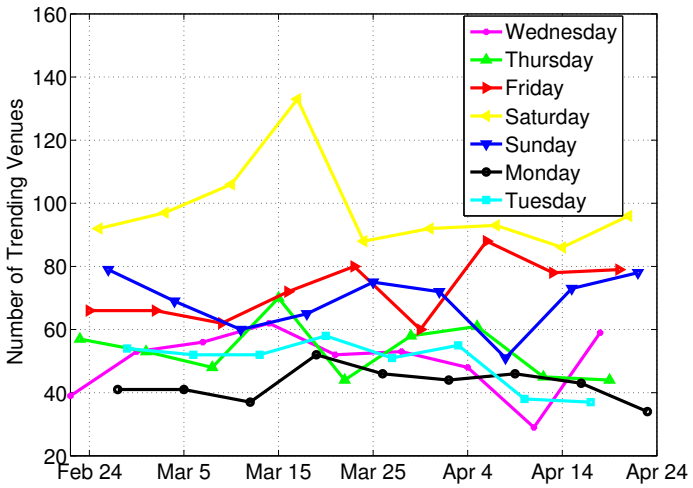

Figure 3.12: Trending venues VS time (day)

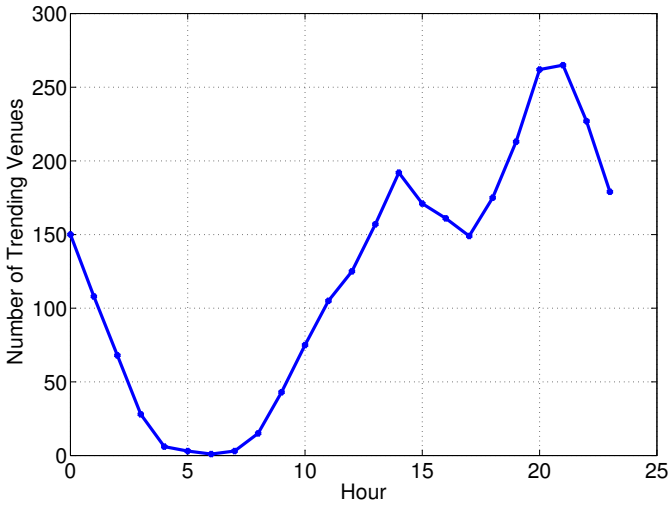

Figure 3.13: Trending venues VS time (hour)

8pm, 10pm and 7pm.

We also count the number of distinct trending venues and the number of their appearances in each category and list them in Figure 3.6. The Food category has the largest number of distinct trending venues and Nightlife Spot category venues appear as trending venues most frequently within the data collection period. None of the Residence category venues appear as a trending venue.

Figure 3.14 plots the CDF of the trending venues' frequency of occurrence. It shows that 26% of the trending venues appear only once. About 50% of the trending venues appear no more than three times. However, about 25% of the trending venues appear at least ten times. These trending venues are very popular and they appear on the trending venue list frequently. For example, Pittsburgh International Airport (PIT) appears as a trending venue almost every hour within our data collection period.

Table 3.6: Statistics of trending venues on categories

Category	Venues	Records
Arts & Entertainment	57	1,845
College & University	57	1,243
Food	183	1,135
Great Outdoors	17	94
Nightlife Spot	149	2,103
Professional & Other Places	69	823
Shop & Service	59	1,537
Travel & Transport	12	1,307

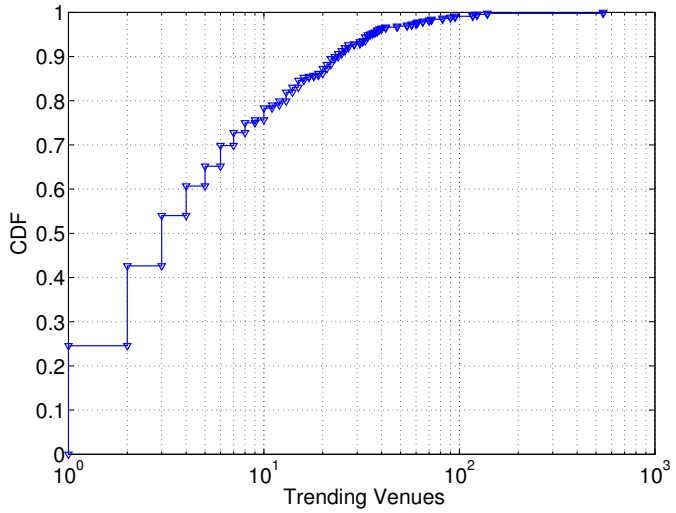

Figure 3.14: CDF of the trending venue appearance

Table 3.7: Summary of the venues with URL, Twitter ID and Menu

	Venues	Trending Venues
URL	12.22%	45.18%
Twitter ID	6.65%	33.13%
Menu	12.52%	36.75%

3.3.2 Relation between Foursquare Features and Temporal Venue Popularity

In Section 3.2.3 we explore the relation between Foursquare features such as specials, web presence and menu and the overall venue popularity. In this subsection, we would like to study the relationship between features and trending venues.

3.3.2.1 Specials In this subsection, we investigate specials. For our data collection period, we get 6,501 specials records in our dataset and there are only 406 trending venue records with such specials (6.25%). However, given that there are only 603 trending venues out of more than 70 thousand venues (<1%), it indicates that there is a correlation between specials and the venues becoming trending venues.

3.3.2.2 Web Presence and Menu We choose the venues in Food category and Nightlife Spot category to investigate the impacts of web presence and menu in making trending venues, because these two categories have the most number of trending venues in our dataset. From these two categories in our dataset, we have 9,723 venues and 332 trending venues. Table 3.7 summarizes the venues and trending venues with Twitter ID, URL and Menu in these two categories. In Food and Nightlife Spot categories, 12.22% of all the venues post URLs and 45.18% of all the trending venues post URLs; 6.65% of all the venues have Twitter IDs and 33.13% of all the trending venues have Twitter IDs; 12.51% of all the venues post Menus and 36.75% of the trending venues have Menus. The higher ratios of trending venues with menu indicate possible influence of the menu feature in promoting a venue to be a

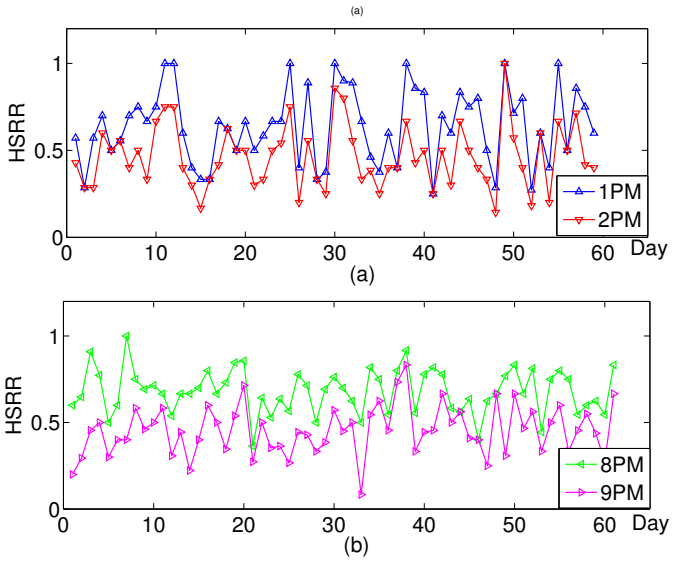

Figure 3.15: Hotness Status Remaning Rates of Trending Venues

trending venue.

3.3.3 Hotness Status

The previous analysis (Figure 3.13) shows that two trending venue peaks at 2pm and 9pm, therefore we choose two two-hour periods (i.e. noon–2pm and 7pm–9pm) from Feb. 23 to Apr. 23 to examine if the trending venues will keep their hotness status in the next two hours. We chose the trending venues at noon and 7pm as the reference venues and we examined how many of these venues are still trending venues in the next two hours. We define the hotness status remaining rate ($HSRR$) between time k and time t ($t > k$) as:

$$HSRR_{k,t} = \frac{||T_t \bigcap T_{t-1} \bigcap ... \bigcap T_k||}{||T_k||} \tag{3.5}$$

Here $a \bigcap b$ stands for the intersection set of a and b. T_t is the trending venues at time t and $||a||$ denotes the number of elements in a. We plot the $HSRR$ in Figure 3.15. Generally, the trending venues can keep their hotness status with high probability. For instance, the

average rate is 64.04% at 1PM and 68.10% at 8PM. After another hour, the average rate is 46.41% at 2PM and 45.46% at 9PM. That is, on average there are more than 60% of trending venues which are able to keep their hotness status each subsequent hour and more than 45% of trending venues remain trending venues after two hours. Moreover, eight trending venues remained as trending venues in the next hour. Therefore, trending venues are able to keep their hotness status during rush hours with high probability.

3.4 DISCUSSION

During the data analysis presented in Section 3.2 and Section 3.3, we observed some interesting findings, which we discuss below.

The Always Popular Venue. We find only one venue which is almost always a trending venue in our dataset–PIT. PIT is also the venue with the largest number of cumulative check-ins and cumulative users in our dataset. In Cheng *et. al*'s work [28], airports are also the most popular venues. We also briefly examined the hometown information of the users checked in at PIT during our crawling period and we found that most users are not from Pittsburgh.

Correlation Between Features And Popular Venues. We investigate the Foursquare features such as specials, web presence and menu to see if there is a correlation between these features and overall and temporal popularity of venues (*i.e.*, trending venues). We do not find strong correlation between specials and attracting more check-ins or users. But there seems a correlation between web presence and menu and venue popularity. When focusing on trending venues, there seems a correlation between all the features and making a venue a trending venue.

Hotness Status of Trending Venues. Our analysis about the hotness status of trending venues shows that trending venues are able to keep their hotness status during a certain time.

3.5 SUMMARY

In this chapter, we have done a comprehensive analysis of the venue popularity about venues in greater Pittsburgh area. We investigate the overall venue popularity, overall category popularity and the trending venue popularity. We also explore the relationship between various features available in Foursquare and venue popularity. Also we study the category popularity based on the predefined category structure. The overall venue popularity analysis are based on the cumulative number of check-ins and the cumulative number of visitors in a specific venue in this chapter. By analyzing venue popularity based on the cumulative dataset in greater Pittsburgh area, we obtain a general view of users' activities regarding local venues on LBSNs. To the best of our knowledge, we are the first to use the cumulative data to investigate the venue popularity on LBSNs. Besides, we study the venue popularity based on the trending venues, which are not cumulative data and collected during a certain time period. Moreover, we briefly look into the spatial features of the trending venues. Therefore, our venue popularity analysis is conducted based on both the cumulative data and the data collected within a certain time period, which demonstrates a more comprehensive overview about the venue popularity. This chapter enables us to understand the users' general preferences of the venues. In the next chapter, we present our study focused on the unobserved group of users' preferences on the venues.

4.0 EXPLORING LATENT GEOGRAPHIC TOPICS

In the previous chapter, we investigated the venue popularity from the entire online population's viewpoint. The overall venue popularity indicates venue reputation based on the entire online populations' preferences. However, sometimes users are more interested in the venues which are popular among similar users (*i.e.*, users with similar tastes and preferences on locations). For example, a freshman may like to know the popular cafes/restaurants among the university students, or a user may be more interested in knowing which restaurants people usually go to after shopping at a mall/watching hockey game. Therefore, it is important to understand the venue popularity based on the certain groups/a certain group of users' preferences on locations.

In this chapter, we would like to study the human preferences of locations based on the movements of unobserved groups of users. That is, we mine the users' check-in records and cluster the venues based on the strong coherence among venues based on by the users' check-ins. The key premise is that the venues that appear together in many users' trajectories will probably be taken as a geographic topic. The venues in the same user mobility-driven geographic topic are potentially and intrinsically related by the human mobility. For example, the university students' check-ins may cause a cluster consisting of campus buildings and so on. In this dissertation we call the clusters of venues mined from the users' check-in records as *latent geographic topics* and we propose a topic-model based approach for venue clustering based on users' historical record of check-ins. Topic models are very common and useful in text mining and in this chapter we present how to mine the latent geographic topics based on the topic model. These mined geographic topics can be used to:

- *Understand a certain group of users' preferences of venues*: The topics are formed based

on users' check-in histories. Thus, the venues in the same geographic topic are the venues that co-exist in check-in histories of a certain number of users. That is, such a topic indicates the preferences of such an unobserved group of users.

- *Recommend venues*: For example, a freshman may be very interested in the most popular cafes among the university students. If there is a geographic topic which includes many University Buildings as well as some restaurants, the cafes in this topic are good venues to be recommended to the freshman.
- *Recommend friends*: since the geographic topics indicate the preferences of the unobserved groups of users, they can also be used in friend recommendation. This is because the users with similar trajectories have higher probability of being friends.

The rest of this chapter is organized as follows. In Section 4.1 we present the summary of the dataset used in this chapter. In Section 4.2 we first introduce the original LDA model in text mining area and the motivation of employing LDA to mining the latent geographic topics of venues. Then we show how we employ LDA in our LBSN data set to obtain the user mobility-driven geographic topics. In Section 4.3, we run the LDA model on our dataset and explain the geographic topics we obtained by using the LDA model. We also investigate the geographic topics based on different temporal aspects, *i.e.*, on weekdays and weekends. In Section 4.4, we focus on travelers and explore their geographic topics. In Section 4.5 we discuss the applications which could use the result of latent geographic topics. In Section 4.6 we explore further of using SVD-based matrix factorization to study the check-in dataset. Lastly, we summarize this chapter in Section 4.7.

4.1 CHECK-IN DATASET

In this chapter, we use the check-in data related to Pittsburgh area collected from *Foursquare* to explore the user mobility-driven local geographic topics. This dataset is different from the cumulative dataset used in Chapter 3, since we use check-ins in this chapter. The Foursquare check-in dataset is collected from Feb 24 to May 23, 2012. We have removed the venues with only one check-in because such single check-ins are not useful for topic mining

	Check-ins	**Venues**	**Users**
All	813,221	16,461	32,113
Weekdays	574,372	16,222	26,224
Weekends	238,849	13,780	22,868

Table 4.1: Summary of the check-in dataset

and may introduce noise in the LDA model that we use. Our dataset used in this chapter is summarized in Table 4.1.

Foursquare defines a hierarchical structure of categories and there are 9 top categories. However, in our data set, there are 45,125 check-ins at the venues that do not belong to any category. Figure 4.1 shows the venue distribution of the other 768,096 check-ins in the top 9 categories. From Figure 4.1, we can see that the number of check-ins in Food and Shop & Service categories are always the largest for weekdays, weekends as well as overall. However, the check-ins in College & University and Professional & Others categories on weekdays are far more than those at weekends. This is not difficult to understand as users usually do not work or study at weekends. The category information can be used to help explain the local geographic topics as described in Section 4.3 and Section 4.4.

4.2 GEOGRAPHIC TOPIC MODELING

In this section, we first introduce the LDA model, which is typically used in text mining area to find the latent topics of texts. Then we show our motivation and how to adopt this model in LBSN area to discover the latent geographic topics.

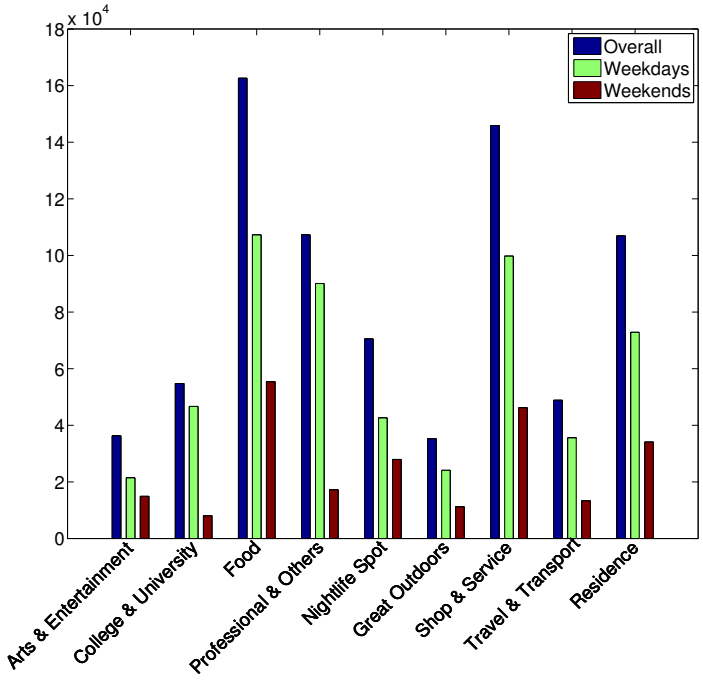

Figure 4.1: Check-ins distribution in top 9 categories

4.2.1 LDA Model

Blei *et al.* present a probabilistic graphical model–Latent Dirichlet Allocation (LDA) in [1]. LDA is usually used to cluster documents based on the topics contained in a corpus of documents. Generally, LDA is able to generate topics of words based on the document inputs and the documents can be described by the generated topics. Thus, before reading a document, we can have a brief idea of the document since we know what the topics are about based on the words of the topics. In Figure 4.2 we briefly show the graphical model of LDA. α and η are parameters of the Dirichlet prior on the per-document topic distributions and per-topic word distributions, respectively. θ_i is the topic distribution for document i. β_k is the word distribution for topic k. w_{ij} is the jth word in document i and z_{ij} is the topic for w_{ij}. In LDA, a word w is the basic unit of data, and there are in total M documents. Each document can be viewed as a mixture of topics z and each topic is consists of words associated to the probability $p(w|z)$. For each document, the probability of a word

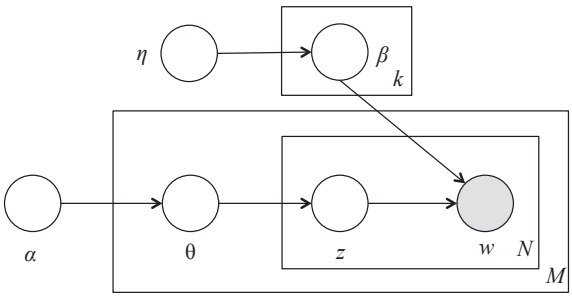

Figure 4.2: Graphical model representation of LDA [1]. The boxes are "plates" representing replicates. The plate M represents documents, while the plate N represents the total words in all documents and k is the number of topics

w_{ij} is given by $p(w_{ij}) = \sum_{t=1}^{k} p(w_{ij}|z_{it})p(z_{it})$, here k is number of topics which can be defined manually. LDA assumes $p(w_{ij}|z_{it})$ and $p(z_{it})$ have multinomial distributions. Once the model is trained, the topics best describing the documents can be extracted (based on $p(d|z)$) and we can present each topic by listing the top words (ranked by $p(w|z)$).

4.2.2 User Mobility-Driven Geographic Topic Modeling

Our motivation of employing LDA to mine the latent geographic topics in LBSNs mainly comes from two aspects. On one hand, comparing to *Latent Semantic Indexing* (*LSI*) [67] which is based on *Singular Value Decomposition* (*SVD*), LDA could also be viewed as a dimensionality reduction technique, but it demonstrates proper underlying generative probabilistic semantics that make sense for the type of data that it models [1]. Moreover, LDA also aims to achieve an illustrative way in which probabilistic models can be scaled up to provide useful inferential machinery in domains involving multiple levels of structure, and it advanced LSI as it can be embedded in a more complex model [1]. On the other hand, before our work, LDA is successfully applied to location related areas in [31, 32, 33, 36, 37, 38]. Ferrari and Mamei present two reasons for choosing the LDA in analyzing users' mobility patterns and routine behaviors in [32]. One is that it is an unsupervised pattern discovering

technique so it does not require labeled data and the other is that the topic results represent meaningful probabilistic distributions over words and documents [32]. Thus, it is possible to adopt the LDA model to identify the geographic topics in our LBSN dataset. However, our work in this chapter is different with theirs. In [31, 32, 33], the focus is on discovering the time-based topics of locations and in [36, 37] the work aims to explore the regional topics. In [38], the study concentrates on analyzing the check-ins and topic model is only a small part of it. In this chapter, we mine the latent geographic topics of venues based on users' check-ins, analyze the topics mined from two different temporal aspects (weekdays and weekends). Besides, we also explore the topics formed by travelers' check-ins.

Next, we explain how we adopt LDA in our LBSN dataset. The basic unit is word in the LDA and, in this dissertation, the venue in a single check-in record represents a word. A user's check-in record consisting of all the venues of his check-ins (check-in history) represents a document, which is a set of words. In this chapter, we focus on the local geographic topics; so M users' check-in histories within a specific city make up the corpus used in the model. Each user's check-in record can be described as a mixture of the geographic topics that are essentially distributions over the geo-locations/venues. A word w which is the basic unit of data represents a venue in a check-in. A set of N words defines the check-in record of the user (*i.e.*, a document). Since there are M users in the dataset, there are M documents. Each user's check-in record can be viewed as a mixture of latent geographic topics z and these topics are distributions over words. k is the number of latent topics. Therefore, the LDA model can be applied to LBSN dataset by using the aforementioned mapping mechanism and an advantage of using LDA is that we do not need to predefine the topics and only need to set the number of the topics.

4.3 EXPERIMENT

In this section, we first describe how we process the dataset to make it suitable for the LDA model and then run the model to get the geographic topics. After that, we present several geographic topics resulting from the experiments. In particular, we first investigate

the overall local geographic topics in our dataset. Since the users' check-in patterns differ from weekdays to weekends [68, 31], we also explore the geographic topics on weekdays and at weekends, respectively. We use the MATLAB Topic Modeling Toolbox to run all our experiments [69].

4.3.1 Data Preparation

In each check-in record, we have the user who created the check-in and the venue where the check-in was created. Moreover, we can get the creation timestamp of the check-in. We conduct three experiments in this section. In the first experiment, we do not consider the creation timestamps of the check-ins. In the second and third experiments, we divide our data set into two subsets according to the creation timestamps of the check-ins, *i.e.*, the weekday subset consists of the check-ins created on weekdays and the weekend subset consists of the check-ins created at weekends.

In these three experiments, the basic units in the user mobility-driven geographic topic model are the venues of the check-ins. A user's check-in history record consists of the venues from a user's check-ins, *e.g.*, $(venue_{check-in1},...,venue_{check-inN})$. We set the number of topics $k = 30$, $\alpha = 50/k$ and $\eta = 0.01$ in all experiments in this section.

4.3.2 Local Geographic Topics

We present the results of the first experiment in this subsection. We get a total of 30 local geographic topics and use OTopics to denote the topics based on overall dataset in the first experiment. In Table 4.2, we list 8 OTopics derived from the first experiment. In each OTopic, we also list top 10 venues. In addition, we discuss the spatial features of these topics.

4.3.2.1 Overall Local Geographic Topics In Section 4.1, we plot the distribution of check-ins in top 9 categories. From Figure 4.1, we can see that the check-ins in Food category and Shop & Service category are the largest. In the generated OTopics, we find that many of them are related to these two categories. For instance, OTopic 7, 9, 19 and 28 are topics

Table 4.2: Examples of user mobility-driven overall local geographic topics

Topic 7	**Topic 9**	**Topic 19**	**Topic 28**
South Hills Village Mall	AMC Loews Waterfront 22	Galleria at Pittsburgh Mills	Walmart Supercenter
Giant Eagle Market District	P.F. Chang's	Cinemark IMAX Theater	Quaker Steak & Lube
Starbucks	Target	Walmart Supercenter	Buffalo Wild Wings
Red Robin Gourmet Burgers	Giant Eagle	Do Drop Inn	Primanti Brothers
Houlihan's Mt. Lebanon	Planet Fitness	Giant Eagle	Costco
T.G.I Friday's	Red Robin Gourmet Burgers	Walmart Supercenter	Giant Eagle Market District
Trader Joe's	Eat'n Park	Target	Giant Eagle
Giant Eagle	T.G.I. Friday's	Applebee's	Starbucks
Walmart	Costco	Giant Eagle	North Park Lounge in HD
GetGo	The Waterfront	UPMC St. Margaret Hospital	Target
Topic 3	**Topic 5**	**Topic 4**	**Topic 11**
Cathedral of Learning	University Center	CONSOL Energy Center	BNY Mellon Center
Hillman Library	Gates-Hillman Complex	PNC Park	Comcast
Hemingway's Cafe	USX Tower	PNC Park	American Eagle HQ
Benedum Hall	Wean Hall	Olive Garden	Fort Pitt Tunnel
Posvar Hall	Hunt Library	Starbucks	Fitness 247
Petersen Events Center	Morewood Gardens	Joe's Crab Shack	PNC Firstside Center
William Pitt Union	Starbucks	Verizon Wireless	"Bellevue, PA"
Peter's Pub	Panther Hollow Inn	St. Clair Hospital	Squirrel Hill Tunnel
Schenley Plaza	Doherty Hall	CCAC Milton Hall	Crawford Square Apartments & Townhomes
Chipotle Mexican Grill	Hamburg Hall	U.S. Steel Clairton Works	Element Church 205 North

related to Food and Shopping categories because most of the top 10 venues in these four topics are venues in Food and Shopping categories. It may indicate that users would likely go to restaurants when they go to shopping. The different topics give an overview of clusters of the shopping venues and restaurants people usually check in at together.

We can also see there are OTopics related to higher education. In OTopic 3, most of the top 10 venues are associated with University of Pittsburgh. OTopic 5 represents the similar case with venues associated with Carnegie Mellon University. Both the universities are located at Oakland neighborhood in Pittsburgh. The food venues in these two topics indicate places that students, faculties and university staffs likely visit frequently.

Some OTopics are related to sports. OTopic 4 is a case in point. CONSOL Energy Center is the home stadium of the hockey team–the Pittsburgh Penguins and PNC Park is the home stadium of the baseball team–the Pittsburgh Pirates. There is also another very famous football team in Pittsburgh–the Pittsburgh Steelers, whose home stadium is Heinz Field. However, it is not in OTopic4 because our data collection period does not overlap with the football season. Thus, the absence of the football stadium confirms that local geographic topics are generated based on the user mobility behaviors rather than the category information.

There are also OTopics that are related to businesses such as OTopic 11, since there are several professional buildings out of the top 10 venues in this topic. An interesting observation in this topic is that there are two tunnels in this topic, which are located in the major route of Pittsburgh (I–376). The existence of these two venues indicates that many users commute between work and home through these two tunnels.

In summary, the advantages of the user mobility-driven local geographic topics include the following:

- the topics generated depend only on the users' check-ins but not on the physical locations or the pre-defined categories.
- the topics provide a novel view of the location clustering based on the human mobility. That is, the venues in a geographic topic imply that people usually go there together with high probability.

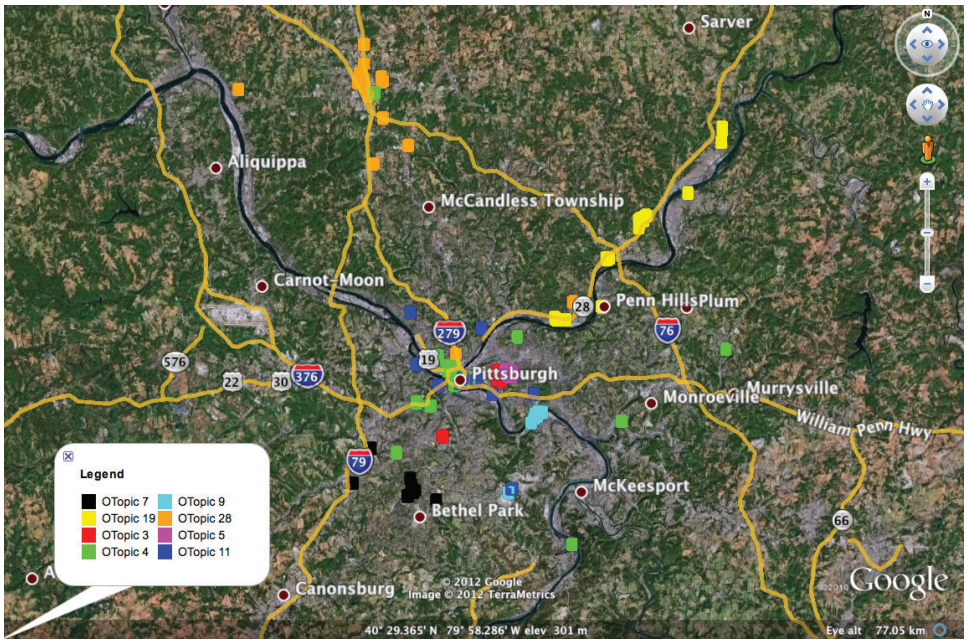

Figure 4.3: The spatial features of the topics

4.3.2.2 Spatial Features of the Topics The spatial features of the generated local geographic topics especially that of the top venues in the topics, are another very interesting issue. We investigate this issue in this subsection, *i.e.*, whether the venues in each topic are geographically close to each other or not? Whether there are any spatial relations among the venues in the same topic?

In Figure 4.3, we show the spatial distribution of the top 10 venues in the eight OTopics illustrated in Table 4.2. We can see that the venues in some OTopics are geographically close, *e.g.*, the top 10 venues in OTopic 5 which are associated with Carnegie Mellon University. The venues in some other OTopics are not very close, *e.g.*, the top 10 venues in OTopic 4. The reason is probably that this is a sports related topic so people may drive to different venues after the the game or drive from different venues before the game. Considering the category information of the topics, venues in the topics related to education (*e.g.*, OTopic 3 and OTopic 5) are usually geographically closer. Venues in topics related to businesses or entertainment may not be close to each other. It is not out of our expectation as people

usually commute between home and work (*e.g.*, OTopic 11) and the stadiums are usually not close to the fans' home (*e.g.*, OTopic 4). However, the case is different for the OTopics related to shopping. Some people would like to go shopping and dining at venues that are not far away. For example, OTopic 9 and OTopic 7; the top 10 venues in these two topics are very close geographically. Some people would like to go shopping and dining at venues that are not that close, *e.g.*, OTopic 19; the top 10 venues in this topic are not close but they are along the freeway.

Figure 4.3 essentially indicates that venues in user mobility-driven local geographic topics are not always geographically close to each other. Our work in this section is a strong complement to the recent work based on the topic model or clustering of the venues in LBSNs presented in [68, 31]. Moreover, our work is useful for business planning. For example, if there is a store or restaurant that is not close to the majority of top venues in a topic, the owner of the store or restaurant may need to consider if a branch office should be opened in the area with the majority of venues in the topic.

4.3.3 Local Geographic Topics on Weekdays vs. Those at Weekends

Users' mobility patterns should be different at weekdays and on weekends. Generally, it is expected that the users' mobility pattern at weekdays should be routine (*e.g.*, commuting between work and home). At weekends, the users' mobility pattern should be diverse (*e.g.*, various entertainment activities and so on). Thus, in this subsection, we present the two experimental results based on different check-ins on weekdays and at weekends and also compare the differences of the geographic topics .

4.3.3.1 Local Geographic Topics on Weekdays We expect that there would be local geographic topics related to universities, professional work and businesses on weekdays and we use WDTopics to denote the topics on weekdays. The results confirm our expectation. We list some topics in Table 4.3. WDTopic 25 relates to University of Pittsburgh and WDTopic 26 relates to Carnegie Mellon University. The top 10 venues in these two topics are almost the same with those in OTopic 3 and OTopic 5 in Section 4.3.2.

Table 4.3: Examples of user mobility-driven geographic topics on weekdays

WDTopic 25	**WDTopic 26**	**WDTopic 5**	**WDTopic 17**
Cathedral of Learning	University Center	UPMC Presbyterian Hospital	Rivers Casino
Hillman Library	Gates-Hillman Complex	UPMC Shadyside	IKEA
Benedum Hall	Wean Hall	Allegheny General Hospital	American Eagle HQ
Hemingway's Cafe	Hunt Library	UPMC Mercy	Comcast
Posvar Hall	BNY Mellon Client Service Center	UPMC Montefiore	Emerald Gardens Apartments
William Pitt Union	Morewood Gardens	UPMC St. Margaret Hospital	Oakmont Tavern
David Lawrence Hall	Porter Hall	Verizon Wireless	Fort Pitt Tunnel
Schenley Plaza	Doherty Hall	Western Pennsylvania Hospital	Rivertowne North Shore
Petersen Events Center	Starbucks	Forbes Regional Hospital	HSLS Falk Library: 200 Scaife Hall
Mad Mex	Tepper School of Business	"Plum, PA"	Squirrel Hill Tunnel
WDTopic 1	**WDTopic 18**	**WDTopic 7**	**WDTopic 9**
CONSOL Energy Center	PNC Park	Pittsburgh International Airport (PIT)	AMC Loews Waterfront 22
Urban Active Fitness	Stage AE	Pittsburgh International Arprt	P.F. Chang's
PNC YMCA	Heinz Hall	The Westin Convention Center	Walmart Supercenter
PITT School Of Information Sciences	Forbes Tower	David L. Lawrence Convention Center	Target
Buffalo Wild Wings	Hamburg Hall	Wyndham Grand Pittsburgh Downtown	Century III Mall
DoubleTree - Green Tree	Heinz Field	T.G.I. Friday's	Giant Eagle
Oakland	PNC Park	Freedom High School	Red Robin Gourmet Burgers
Bear Run Village	Olive Garden	Heritage Hills Townhomes Apartments	Eat'n Park
Renaissance Pittsburgh Hotel	Starbucks	EFI, Inc. (Pittsburgh Office)	Amberson Towers
Ariba Inc.	Pittsburgh International Airport (PIT)	Tonic Bar And Grill	Jefferson Regional Medical Center

We can see there are several WDTopics related to professional work and business. For instance, WDTopic 5 is related to medical work, as 8 out of 10 top venues in this topic are hospitals. We consider two possible reasons for forming this topic. One possible reason could be that patients usually go to these hospitals for different purposes. This is because University of Pittsburgh Medical Center (UPMC) is the largest medical center in western Pennsylvania area, so patients may have been referred to different specialists who work in different hospitals belonging to the UPMC. The other reason could be that the doctors, nurses and other medical staff work at different locations at different times. There is no topic related to the medical work in the overall set of topics.

We also have a topic (WDTopic 17) with two tunnels–Fort Pitt Tunnel and Squirrel Hill Tunnel that are also included in OTopic 11 in Section 4.3.2. However, the top 10 venues in WDTopic 17 are a little different from those in OTopic 11. Rivers Casino and IKEA are in WDTopic 17 but not in OTopic 11. The reason may be that there are many people who work at Rivers Casino and IKEA and usually commute on the main route I–376 on weekdays.

The WDTopics related to sports are also a little different from the OTopic related to sports. We can see that CONSOL Energy Center and PNC Park are in two topics (one in WDTopic 1 and the other is in WDTopic 18). In both topics the hotel venues in the top 10 venues may imply that hockey fans or baseball fans go to Pittsburgh for a game on weekdays. We can see University of Pittsburgh School of Information Sciences and Oakland (University of Pittsburgh and Carnegie Mellon University are both at Oakland) in WDTopic 1, which indicates that university students are interested in watching the early hockey game on weekdays through getting student discounts [70].

WDTopic 7 shows that the airport on weekdays are related to businesses, as there are many conferences or events at the convention centers such as Westin Convention Center, David L. Lawrence Convention Center and Wyndham Grand Pittsburgh Downtown.

WDTopic 9 gives an example of the topics in weekdays related to shopping and food. The venues in WDTopic 9 are almost the same as the venues in OTopic 9 in Section 4.3.2.

4.3.3.2 Local Geographic Topics at Weekends

At weekends, we expect the topics are more related to entertainment and shopping activities. Besides, there would be very few

Table 4.4: Examples of user mobility-driven geographic topics at weekends

WETopic 4	**WETopic 16**	**WETopic 10**	**WETopic 27**
Pittsburgh International Airport (PIT)	Heinz Hall	Target	The Cheesecake Factory
Pittsburgh International Arprt	Carnegie Science Center	Eat'n Park	Bar Louie
Hard Rock Cafe Pittsburgh	Pittsburgh Zoo & PPG Aquarium	Giant Eagle	Claddagh Irish Pub
Robert Morris University Island Sports Center	Petersen Events Center	The Waterfront	Pittsburgh Marriott City Center
T.G.I. Friday's	Red Robin Gourmet Burgers	Dave & Buster's	Emerald Gardens Apartments
3:36	Carnegie Museum of Natural History	T.G.I. Friday's	The Altar Bar
Baggage Claim	Joe's Crab Shack	AMC Loews Waterfront 22	Grand Concourse
Hard Rock Cafe	Station Square	Barnes & Noble	Megabus Pittsburgh
"Freedom, Pa"	Bar Room	Costco	Starbucks
House	Eat'n Park	Giant Eagle	Joe's Crab Shack

topics related to education, business and professional work. Our results indeed confirm our expectation. There is no topic related to any university in the 30 topics, which indicates that there are not as many users at weekends checking in at universities as those on weekdays in *Foursquare*. There is no topic about professional work or business, either. Almost all the topics are related to food, shopping and entertainment. We use WETopics to denote the local geographic topics at weekends and Table 4.4 lists some examples.

WETopic 4 shows the geographic topic related to airport. We do not see any convention centers in the top 10 venues in this topic, which may imply that users usually do not take flight for business reasons at weekends. WETopic 16 is mainly related to arts and entertainment, as the top 10 venues are associated with such categories. Many concerts, operas and stage shows are held in Heinz Hall. Carnegie Science Center and Carnegie Museum of Natural History are very famous museums in Pittsburgh. Petersen Center is the sport center of University of Pittsburgh, which usually hosts a lot of university sport events. WETopic 10 is related to shopping and food. The top 10 venues in this topic are very similar with those in OTopic 9 in Section 4.3.2. WETopic 25 is mainly related to food, which is different from the OTopics related to food, in which food venues usually appear together with the venues in shopping or the venues in entertainment.

4.3.3.3 Comparisons Between WDTopics and WETopics

We compare all the local geographic topics on weekdays and those at weekends and we observe the following:

- We do not find any topic related to education, business or professional work out of all 30 WETopics. The reason for this maybe that people usually do not go to school or work at weekends and they rather would like to engage in social, family events and recreations at weekends. Thus, the topics at weekends are related to entertainment and recreation activities. We also observe that the number of topics related to shopping and various entertainment is more at weekends than that on weekdays.
- There are topics related to only food category at weekends. However, there is no such topics on weekdays. On weekdays, Food venues usually appear together with shopping, entertainment or business venues in the local geographic topics.
- On weekdays, there are only a few art and science center venues that appear in the

geographic topics, however, there are many such venues in several geographic topics at weekends. One possible reason could be that these art and science centers are usually closed after 5pm on weekdays, thus people may have no time to visit such centers after the work (Figure 3.13 indicates the main check-in peaks are after 5pm). Another possible reason could be that people would like to visit them on weekends as it usually takes long time to visit such museums and art centers.

In summary, the differences of the human mobility behavior on weekdays and at weekends determine the differences between the local geographic topics on weekdays and those at weekends.

4.4 EXPLORE THE TRAVELERS' GEOGRAPHIC TOPICS

Intrinsically, travelers' activities in LBSN are distinctive, when compared with local users' activities. For instance, travelers usually prefer to visit the famous venues or landmarks. On the contrary, local users' check-ins may be more routine compared to the travelers' check-ins. Therefore, a study of travelers' activities in LBSNs is important to understand traveler behavior and also can help LBSNs provider to improve their services, *e.g.*, location recommendation service to travelers/new visitors, which is especially crucial for a new visitor in a city. Thus, in this section, we focus on studying the check-ins of travelers in a specific city. In particular, we first empirically investigate the temporal and spatial features of travelers' check-ins. Then we employ the LDA model to mine the latent geographic topics from travelers' check-ins created in the greater Pittsburgh area. We believe our work can better help to understand the travelers' requirements and then improve the location recommendations to them by analyzing latent geographic topics by the travelers.

4.4.1 Data Preparation

In Foursquare, users can create profiles and add their names, photos, hometowns, *etc.*, to the profiles. We define a traveler as a user whose hometown is more than about 310 miles away

Figure 4.4: Word cloud based on travelers' check-ins

from downtown Pittsburgh. Moreover, we have removed the venues with very few check-ins as such check-ins may introduce noise in the LDA model that we use. We keep monitoring the check-ins until July, 2012 and we extract the travelers' check-ins. We finally have 104,887 check-ins created on 8,016 venues. These check-ins were created by 8,746 travelers. We show the word cloud based on the check-ins of this data set in Figure 4.4[1]. The size of each word is proportional to the probability of its appearances in the data set. It is obvious that the PIT Airport[2] is the most visited venue by travelers[3]. King and Pozdnoukhov also report the airports are very popular venues in [71]. Moreover, traveling by air is one of the main transportations so it is not surprising that PIT Airport is the most popular venue.

4.4.2 General Analysis

We plot the distribution of the venues and check-ins in the top 9 categories in Figure 4.5 and Figure 4.6. Figure 4.5 shows that the top 3 categories based on the number of venues are Food, Shop & Service and Professional & Other Places categories. With regards to travelers'

[1]The figure is generated with http://www.wordle.net/.

[2]The full name of PIT Airport is Pittsburgh International Airport (PIT). However, the name is too long to show in the figure, so we use a short name instead.

[3]Pittsburgh International Arprt is the same venue as the PIT Airport, however, their venueID are different, so we do not merge these two venues into one.

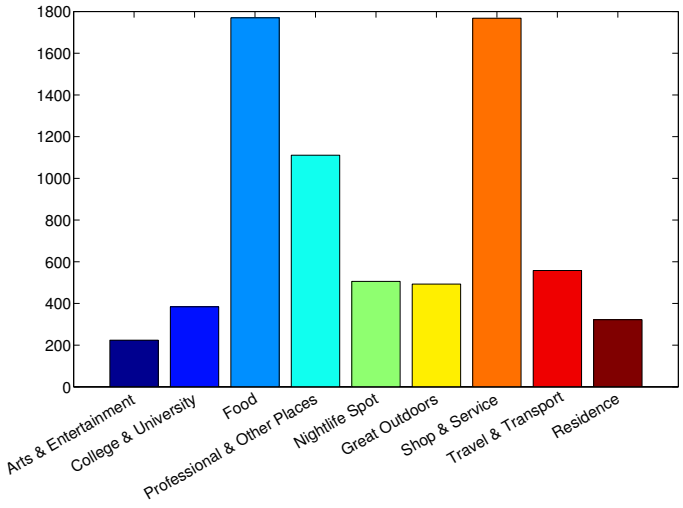

Figure 4.5: Venue distribution in top 9 categories of travelers' check-in dataset

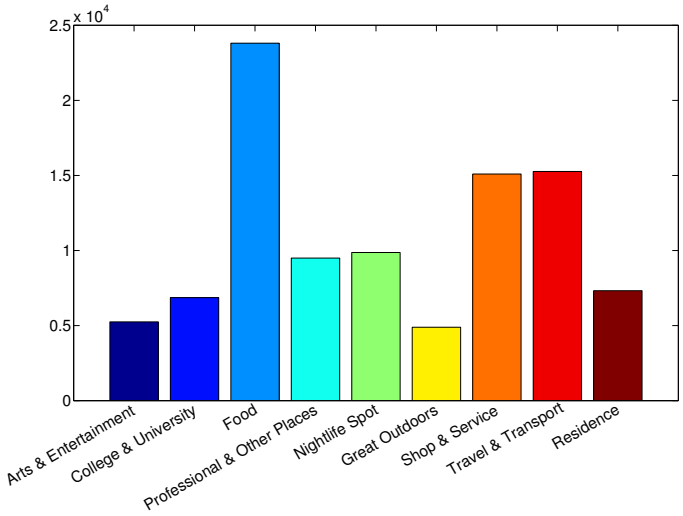

Figure 4.6: Check-in distribution in 9 top categories of travelers' check-in dataset

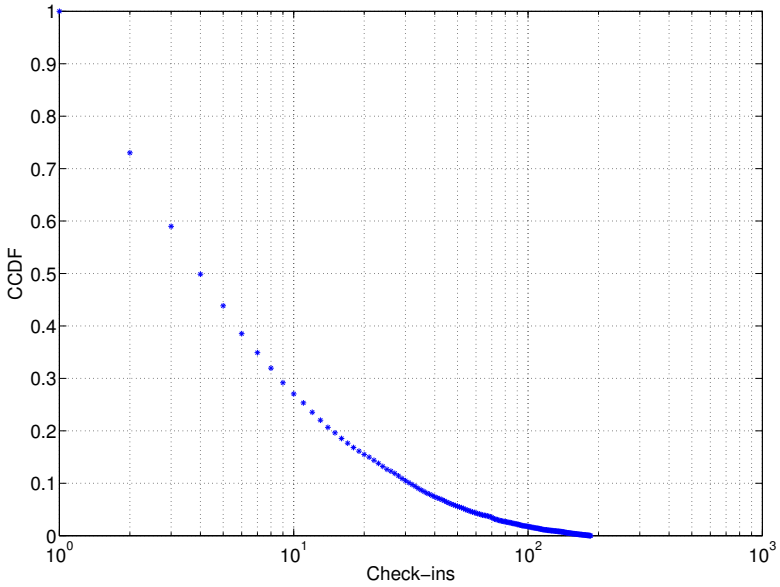

Figure 4.7: CCDF of the total check-ins of each venue

check-ins shown in Figure 4.6, venues in Food category still attract the largest number of check-ins. However, check-ins created at venues in Travel & Transport category are more than that in Shop & Service category. Since the Food category and Shop & Service category have more venues than others, it is not surprising that these two categories attract huge number of check-ins. But comparing these two categories, we find that travelers prefer to check in at venues in Food category other than those in Shop & Service category, as these two categories have almost the same number of venues but there are much more check-ins in Food categories.

4.4.3 Analysis of Check-ins

From the 8,016 venues, we plot the CCDF of the total check-ins of each venue in Figure 4.7. It is obvious that 50% of the venues have no more than 4 check-ins and less than 10% of the venues have more than 12 check-ins. However, there are some very popular venues (about 3%) which attract more than 100 check-ins.

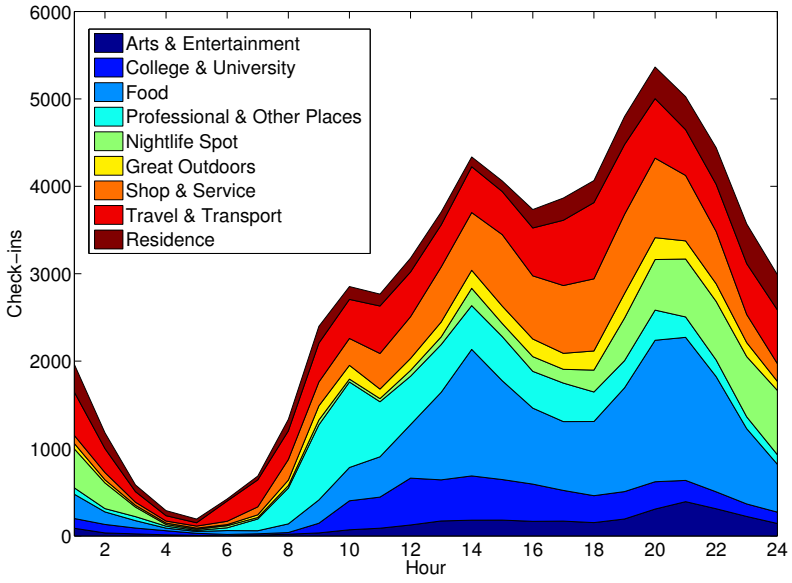

Figure 4.8: Check-ins in each category on weekdays

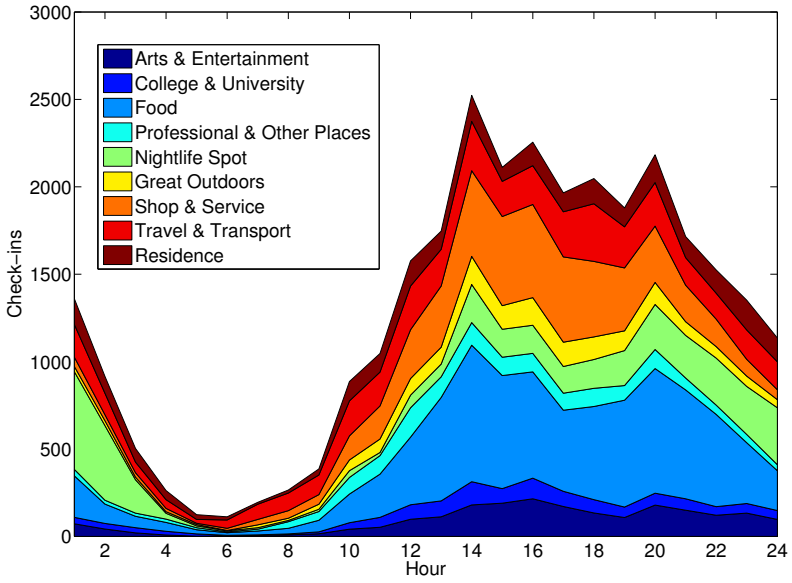

Figure 4.9: Check-ins in each category at weekends

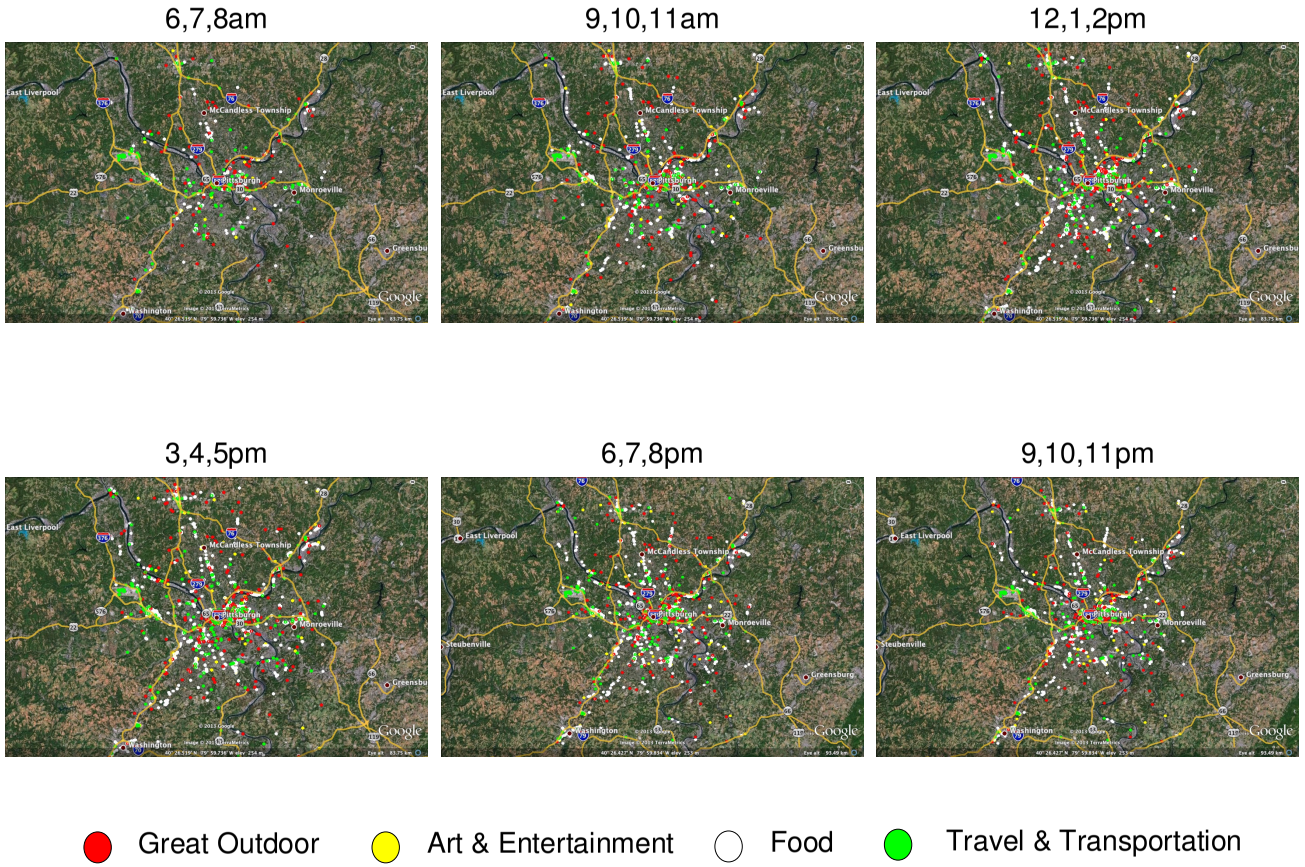

Figure 4.10: The evolution of the venues in Great Outdoor, Art & Entertainment, Food and Travel & Transportation categories based on spatial-temporal information

We also study the temporal feature of the check-ins for each category. We plot the distribution of the check-ins on weekdays in each category by hour in Figure 4.8 and plot the distribution of the check-ins at weekends in each category by hour in Figure 4.9. From these two figures, we can see that the check-in peak on weekdays is at 8pm but the peak at weekends is 2pm. We can also find that travelers usually visit Art & Entertainment venues at weekends. Moreover, we find the check-ins exhibit a general rising trend from 5am until 8pm on weekdays and then drop quickly after 8pm. However, travelers' check-in trends are different at weekends. The check-ins increase rapidly and peak at 2pm; after that, the check-ins show a very slow decrease from 2pm to 8pm. These two different trends indicate that travelers usually hang out during the afternoon at weekends but on weekdays travelers' activities spread out through the day.

Moreover, we investigate the spatial-temporal evolution of the venues that users have checked in at in Great Outdoor category, Art & Entertainment category, Food category and Travel & Transportation category. We use a three-hour time slot as the basic temporal unit to study the evolution of travelers' check-ins. We show the result in Figure 4.10. We can see that generally the Pittsburgh Downtown area (the center of the map) is always a hot area and the check-ins gradually spread out from the Downtown area over time. Such a trend is especially remarkable with regards to the venues in Food category and Great Outdoor category.

4.4.4 Latent Geographic Topics

In this subsection, we mine the traveler dataset using the LDA model to investigate the latent geographic topics of the travelers' check-ins. We set the latent topic number $k = 10$, the parameter of the Dirichlet prior on the per-document topic distributions $\alpha = 50/k$ and the parameter of per-topic word distributions $\eta = 0.01$ in the experiment in this subsection. We present some of the interesting topics that we mined through the LDA model as follows.

4.4.4.1 Topics related to Sports Pittsburgh is a well-known sports city. Therefore, we can expect that a lot of sports fans would visit Pittsburgh for watching games. The

Table 4.5: Topic related to hockey

Venue	**Category**
Pittsburgh International Airport (PIT)	Airport
CONSOL Energy Center	Hockey Arena
Pittsburgh Marriott City Center	Hotel
DoubleTree Hotel	Hotel
Linzer Apts	Apartment
Baggage Claim	Airport
Home Depot	Store
F'n Nottingham	Uncategorized
Fox & Hound English Pub & Grille	Bar
The Steel Mill	Residence

geographic topics related to the sports confirm it. Table 4.5 lists the top 10 venues of a topic related to hockey. The top 2 venues in this topic are Airport and the hockey arena. Besides, there are two hotels in this topic, which implies the travelers may stay at these hotels. Thus, we can conclude that watching games is one of the main goals of the travelers to Pittsburgh.

4.4.4.2 Topics related to Higher Education University of Pittsburgh and Carnegie Mellon University (CMU) are two very famous universities in Pittsburgh. We anticipate that travelers will visit these two universities for academic reasons (academic conferences, seminars and so on) or family reasons (parents visit their children who study at the university). One latent geographic topic verifies it and we list the top 10 venues of this topic in Table 4.6. In this table, we can see the majority of the venues are CMU buildings. The bar and the coffee store indicate the preferences of the travelers who usually visit CMU.

Table 4.6: Topic related to high education

Venue	**Category**
Gates-Hillman Complex	CMU
University Center	CMU
Mario's East Side Saloon	Bar
Hamburg Hall	CMU
Hunt Library	CMU
Heinz Hall	Concert Hall
Starbucks	Coffee
Union Grill	Restaurant
Tepper School of Business	CMU
Wean Hall	CMU

4.4.4.3 Topics related to Transportation & Hotels We also observe a topic related to transportation and hotels, which is not out of our expectation. Table 4.7 shows this topic and we can see that most of the top 10 venues are airports and hotels/lodges. The transportation and hotels/lodges are very important to a traveler's trip. Convenient transportation and comfortable hotels/lodges could greatly improve the satisfaction of the travel. This geographic topic implies that many travelers choose airplanes for their travel and many travelers choose the hotels in this topic as the place to stay in Pittsburgh. Moreover, many users who check in at both the airports and the hotels in this topic may visit the bar in the same topic. Therefore, this topic is helpful in hotel or restaurant recommendations to the travelers.

Table 4.7: Topic related to Transportation & Hotels

Venue	**Category**
Pittsburgh International Airport (PIT)	Airport
Pittsburgh International Arprt	Airport
The Westin Convention Center	Hotel
Ho-Jo's	Uncategorized
Hyatt Regency Pittsburgh International Airport	Airport
The Cat's Pajamas	Residence
Tonic Bar And Grill	Bar
pittsburg airport	Airport
Hamilton Abode	Residence
The Hamiltons	Travel

4.5 APPLICATIONS

We discuss how the proposed latent topic modeling and the latent geographic topics can be used in different applications.

Location Recommendation: Our proposed latent topic model based approach can be very helpful for location recommendation. The latent topic model is based on the users' historical check-ins, thus the venues in the same topic are the ones that many people usually go to. For example, some parents would like to take their children to visit the museums or science centers at weekends. Thus, the topics about entertainment could provide very valuable references since many people like to visit such POIs.

Friend Recommendation Since each user's historical check-in record can be described by the topics, the similarity of the topics could be helpful in recommending friends to users because of the similar interests of POIs (*i.e.*, placefriends [72, 73]). For example, for two PITT students who usually check in at Hillman Library, School of Information Sciences and William Pitt Union (all are PITT buildings) and although they may not know each other

right now, the probability for them being friends are high. This is because they may register in the same class in the same building, or join the same student organization or interest group and meet in the student center, thus the location information is also important in friend recommendation.

Business Strategy Design: Users' mobility-driven local geographic topics also benefit the business strategy design. The geographic topics can help a business owner to find out whether there is any complementary relationship between his venues and other venues and then help him explore the potential locations for a new store. It is also helpful for the business owners to identify his competitors to improve his own business.

4.6 MORE EXPLORATIONS AND DISCUSSION

We show how we use LDA to mine the LBSN check-in dataset to explore the groups of users' preferences of locations in the previous sections. Here, we would like to explore more using other models. In [74], Zheng *et al.* summarize the the typical technologies that can be used to analyze multimodal data including location information and matrix factorization is one of them. Therefore, in this section, we present our exploration of mining the check-in dataset using the SVD-based matrix factorization.

4.6.1 Singular Value Decomposition

SVD is a method to decompose a matrix of interest into a product of matrices:

$$A_{[m\times n]} = U_{[m\times r]}\Sigma_{r\times r}(V_{[n\times r]})^T \tag{4.1}$$

In Equation 4.1, A is an $m \times n$ rank-r matrix. $U(V)$ is an $m \times r$ $(n \times r)$ matrix that includes the left (right) singular vectors of A while Σ is a diagonal $r \times r$ matrix whose diagonal elements are the singular values of A in descending order.

Each check-in record in our check-in dataset contains a user who created this check-in and a venue where the check-in is created, thus the check-in dataset can be presented by

a $m \times n$ matrix A, where m is the number of users and n is the number of venues in our dataset. Therefore, the SVD could decompose A into the product of matrices as shown in Equation 4.1. The left singular vectors U then maps each user to the geo-concept space and right singular vectors V maps each venue to the geo-concept space. The singular values Σ describe the strength of each geo-concept in the dataset. By using only the k-largest singular values we can approximate the original matrix A with a rank-k matrix (by setting the $(r-k)$ smallest singular values in Equation 4.1 to zero). A rule-of-thumb for choosing k is to retain 90-95% of the total energy of the singular values (*i.e.*, $\frac{\sum_{i=1}^{k} \lambda_i^2}{\sum_{i=1}^{r} \lambda_i^2} > 0.9$).

4.6.2 Exploring the check-in dataset by SVD

From Table 4.1 we can see that our user-venue check-in matrix is very sparse, thus we use the MATLAB *svds* function to decompose the check-in matrix. We investigate the right singular vectors V for possible latent features of the check-ins as the singular vectors V implies a mapping from venues to the geo-concept space.

In Section 4.6.1 we introduce the approximation method of keeping the top k biggest singular values from Σ whose summation is larger than a threshold ratio of the total summation of all the singular values. But MATLAB *svds* function provides to compute the k largest singular values and associated singular vectors of matrix A. Therefore, in our experiment we choose to run *svds* based on different $k = 20, 30, 40$ and 50, respectively. Then, we extract the top 10 venues in each row in V^T (by choosing the top 10 maximum abs value in each row).

We list the first geo-concept and third geo-concept in Table 4.8. The first geo-concept corresponds to the largest singular value in Σ. We carefully examine the venue "The Assassin Tower" and we find that venue gets the most number of check-in from an individual user u_i. We then check the check-ins created by the user u_i and we find that the top 5 venues in geo-concept 1 are exactly the top 5 venues visited by the user u_i based on the number of check-ins. We also did the same for the third geo-concept 3 and the similar finding also applies to this geo-concept. That is, a user u_j who created the second largest number of check-ins at the top 1 venue in geo-concept 3 and top 5 venues in geo-concept 3 are also top

Table 4.8: Examples of SVD-based geo-concepts

Geo-concept 1	**Geo-concept 3**
The Assassin Tower	Home
Bakery Cakery Brodega	Missy's House
Uptown mini-mart	Lindsay's House
Moon Jail	Target
Primanti Bros.	Longwood at Oakmont
Morewood Gardens	Grandma's House
Gates-Hillman Complex	Walmart Supercenter
The Library	Mcdonald's
Fisher Hall	Giant Eagle
Giant Eagle	Home

venues visited by the user u_j.

Given the large scale nature of our dataset and its sparsity–*i.e.*, users perform few check-ins on average–matrix A is of hight rank. Hence, the number of latent topics/concepts that SVD will identify to capture 90-95% of the energy/variance in the original dataset is expected to be high–and much higher than the 30 topics we manually set in LDA.

4.6.3 Discussions

LDA and SVD take a different approach on identifying latent patterns in the dataset. Since the ground truth is not known we cannot compare the two models in absolute terms but we can qualitatively compare them with respect to the information revealead by them. In particular, one of the imporant aspects of the two models is the number of topics to be identified.

With LDA the number of topics need to be manually set, while with SVD there is a principled approach on the choice of k. However, as alluded to above, since our dataset

is very sparse, in order to keep the majority of the information in our dataset SVD will "generate" a large number of topics, which might be not practical/convenient in some case. On the other hand with LDA, we are able to manually set the number of concepts we want to identify and hence, we can obtain a small set of topics that potentially describe our dataset. Nevertheless, there is no guarantee that the number we chose is an appropriate one. On top of that, if we want to examine $k' > k$ topics, LDA might provide a totally different set of topics. This is not the case of course with SVD and hence, it can create further interpret-ability issues.

4.7 SUMMARY

In this chapter, we study further in understanding users' preferences of venues through mining the check-in data. In particular, we employ the LDA model to mine the check-in dataset in order to get the latent geographic topics. We explain the LDA model and how to use it in the LBSN check-in dataset to get the user mobility-driven geographic topics. The mined geographic topics is very helpful in understanding certain groups/a certain group of users' preferences of locations. We also investigate the differences of the user mobility pattern on weekdays and weekends. Moreover, we use the hometown information to filter the travelers from the dataset and explore the traveler behavior in this chapter. We analyze the travelers' check-in distribution of the top 9 venue categories and investigate the temporal and spacial feature of the travelers' check-ins. We also employ the LDA model on the travelers' check-in dataset and the obtained geographic topics can help to understand the travelers' behavior. After that, we briefly discuss some applications that can use the LDA model and geographic topics. Lastly, we explore further about mining the latent patterns of the check-ind dataset by SVD-based matrix factorization and discuss our results.

5.0 HYBRID TRUST-BASED POI RECOMMENDATION

In Chapter 3 we studied the entire online population's preferences on locations and in Chapter 4 we explored the latent geographic topics derived from users' check-in records, which indicate the unobserved group of users' location preferences. In this chapter, we mainly focus on the individual user's preferences of locations, *i.e.*, we leverage the individual user's check-in records and his social network to propose a personalized hybrid trust-based POI recommendation approach.

It is well known that existing web recommendation systems usually employ algorithms such as item-based recommendation [75, 76, 77], matrix factorization techniques [78, 79], collaborative filtering (CF)-based recommendation and recommendations based on social links [80, 81, 82, 83]. There are several challenges of directly employing the current web recommendation algorithms for POI recommendation in LBSNs. One challenge is how to integrate geographic information related to the POIs into the recommendation. For example, in a platform such as Netflix or Amazon recommending movies on Netflix to users or recommending books to users do not consider the geographic information into the recommendation algorithm. However, when recommending POIs to users, the geographic information should be considered. For instance, recommending the Disney Theme Park in Florida to a user who is now in Chicago is not a good recommendation. Thus, when making POI recommendations, we should take the distances of the POIs into consideration. Another challenge is that the rating systems are not generally supported by the current LBSNs; this adds difficulties in measuring users' preferences with regards to POIs in LBSNs. For example, when recommending an item to a user in Amazon.com, the ratings of the similar items can be used in recommendation. Moreover, in the platforms such as Amazon or Netflix, or even in web search engines, when recommending the items the social influences is usually not

considered. Social links are usually used in friend recommendation in social networks [84]. In such social network platforms, nodes represent users. Social reasons are crucial for users' check-ins at various locations. For example, if a friend recommends a good restaurant to a user, the user probably visits the restaurant as he would likely trust his friends. Besides, a user may visit his relatives/friends and hang out with friends, so it is necessary to consider the social aspects into the POI recommendation. Hence, employing just the state-of-the-art recommendation approaches that are mainly based on users' preferences of the items for recommendations based on various contexts will not be adequate for POI recommendation [11].

In LBSNs, users visit POIs for various reasons. As mentioned earlier, social reasons play an important role in recommending POIs. User's check-in behavior also plays a part in POI recommendation. Our venue popularity analysis in Section 3.2.1 shows popular venues attract a lot of users, thus we should also consider such a factor into POI recommendation. Although it is not easy to capture such diverse reasons accurately and completely, it is necessary to integrate such reasons into the POI recommendation. In this chapter, we mainly focus on the users' social networks and the check-in behavior and present our proposed personalized hybrid trust-based POI recommendation. The contributions of the proposed approach are as follows.

- The proposed trust-based recommendation provides a novel approach to evaluating the POI reputation based on both users' social networks and users' check-in activities. It is a novel hybrid trust-based model, that integrates graph-based model and interaction based model; and contributes to this important area where there is limited work, as mentioned in [44].

- The proposed approach is flexible as users can change the parameters of the proposed approach to make a personalized recommendation. Besides, the approach supports different types of weight functions, which are easy to tailor to the specific applications. Moreover, although the proposed hybrid trust model is based on building trust through users' check-in activities on POIs, it is applicable to other applications that allow interactions between users and other items.

The rest of this chapter is organized as follows. In Section 5.1, we present the motivation for the proposed model and the problem definition. In Section 5.2, we introduce three models which are closely related to our proposed model. In Section 5.3, we present the proposed hybrid trust-based POI recommendation in detail. In Section 5.4, we present the dataset used in this chapter. In particular, we analyze both the social graph information and check-in information. In Section 5.5, we introduce the methodology and measures used to evaluate the proposed model. We also present the experimental results. In Section 5.6, we discuss the proposed approach and we present the summary in Section 5.7.

5.1 MOTIVATION & PROBLEM DEFINITION

In this section, we present the motivation of our proposed hybrid trust-based POI recommendation approach for LBSNs and the problem definition.

5.1.1 Motivation

From the analysis of the previous chapters, we can see that popular POIs attract large number of users and check-ins, *i.e.*, users usually love to visit POIs with good reputation. Hence, the POI reputations can be used to rank the POIs and we can recommend the POIs based on their reputation to LBSN users. How to measure the POI reputation is thus a crucial challenge. Since the popular POIs are determined by human activities, thus, users' check-in behavior should be considered in POI reputation calculation. Our goal is to provide a personalized POI recommendation, thus we should consider social connections in LBSNs which may impact an individual user's location preferences. For example, if a friend recommends a POI to a LBSN user, the user will visit the POI with higher probability since he likely trusts his friends. Therefore, we propose to compute a POI's reputation based on its visitors' trust values. In this way, we employ both the social information and check-ins available in LBSNs to calculate the POI reputation.

5.1.2 Problem Definition

We present the problem definition in this subsection. For POI recommendation in a LBSN, we consider a set of users $U = \{u_1, u_2, \ldots, u_N\}$ and a set of POIs $P = \{p_1, p_2, \ldots, p_M\}$. A user u_i can be a friend of another user u_j and we define the edge between these two users as $e_{i,j} = (u_i \leftarrow u_j)$ when they are friends of each other. We define the social graph of the LBSN users as $S(U, E_u)$ where E_u is the set of friendship links among the users in user set U, $e_{i,j} = (u_i \leftarrow u_j) \in E_u$. Here $u_i \in U$ links to $u_j \in U$ and $1 \leq i, j \leq N$. In a LBSN, a user u_i's check-in on a POI p_k can be defined as an edge $e_{u_i,p_k} = (u_i \rightarrow p_k)$. Therefore, users' interaction/check-in graph can be defined as $C(U, P, E_c)$ where E_c is the set of check-in actions $e_{u_i,p_k} = (u_i \rightarrow p_k) \in E_c$. Here $u_i \in U$ and $1 \leq i \leq N$, $p_k \in P$ and $1 \leq k \leq M$.

Our key goal is to recommend POIs that have good reputation. Thus, the task of personalized POI recommendation can be described as follows: Given a user u and a POI p, if the reputation of p from the perspective of u is unknown, compute the reputation of p for u. The reputation of p to user u is denoted by r_p (r_p is different for different users). Figure 5.1 illustrates the personalized trust-based POI recommendation problem. As shown in the figure, we know the social graph and available POIs. The check-in histories of the users are also known. User A wants to get a recommendation of a POI which she has not visited. Then the goal of the recommendation system is to find the POI with highest reputation based on the social graph and the historical interactions between the users and POIs.

Although many disciplines have studied trust and each disciple defines trust from its own perspective, a general understanding about trust is a measure of confidence of the expected behavior of an entity or entities [44]. In this dissertation, we define trust as *an individual user's perceived confidence of the authorities of other users' check-in activities, which could be characterized from two aspects–social relationship in a network and the check-ins created by users on POIs.* To a specific user, the other users' trust values are determined by their social relationship with him and the merits of their check-in histories. We will present how to compute the users' trust values and POI reputation in Section 5.3.

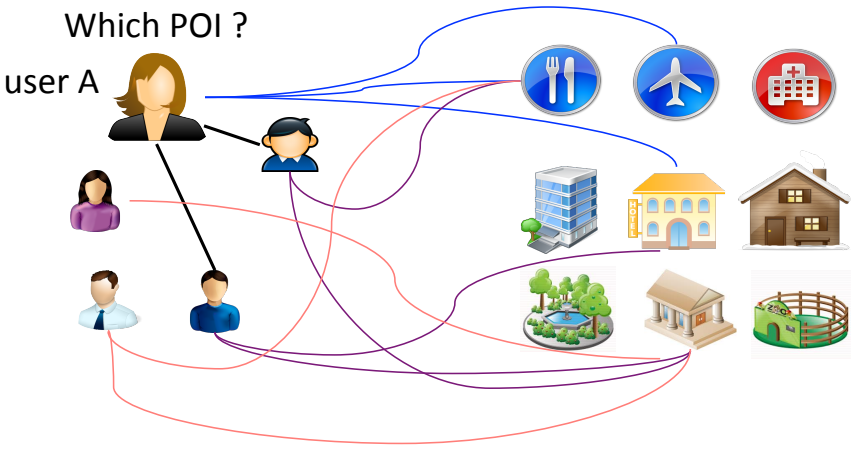

Figure 5.1: Illustration of a LBSN network with users and users' check-ins.

5.2 BACKGROUND

Our proposed hybrid trust-based POI recommendation approach is motivated by graph-based trust model, interaction-based trust model and HITS (Hypertext Induced Topic Search) model [85]. The graph-based trust model helps in addressing the issue of how to use the social links in evaluating users' trust value. The interaction-based trust model motivates using the check-in activities created by users on POIs in evaluating users' trust value. Although there is only one kind of node in HITS model, the concepts of authorities and hubs inspires us to combine the graph-based trust model and interaction-based trust model into the proposed trust-based recommendation approach to calculate POIs' reputations. We present the three models in this subsection.

5.2.1 Graph-Based Trust Model

Generally, the graph-based trust model evaluates the users' trust value based on the structure of the social graph. For example, the concept such as Friend-Of-A-Friend (FOAF) is widely used in graph-based trust model, which exploits the propagation property of the trust to

compute trust among users in social network [44]. For example, if the social network is an unweighted graph, to a user, the trust values of his friends are higher than those of non-friend users. Also, trust will propagate along the friendship links. Thus, a friend's trust value is higher than the trust of the friends of this friend. Moreover, the social network can be a weighted graph. For example, the weights of edges connecting the family members of a user on the social network can be set higher than that of edges connecting with friends. The trust value can be also propagated through such a weighted graph. There are also other examples such as including the feedback to evaluate trust value [45]. All such approaches leverage network structures to calculate user trust value, and they capture the trust on how users are related to each other and how trust flows through their network [44].

LBSNs are also social networks and users can create friendship links with other users. Besides, when a friend recommends a POI to a user, the user probably will visit the POI because he trusts his friend. Thus, the social graph of the LBSN is helpful in the POI recommendation and we consider the graph-based trust model in our proposed hybrid model.

5.2.2 Interaction-Based Trust Model

Besides the graph-based trust model, the interactions among the users can be also used in computing the trust values and such a model is called as the interaction-based trust model. The interactions can be the reviews, comments or ratings in the online community [52], or the acceptance and approval of a member in the community and the engagement/involvement of the member in the community [53], or the frequency and the duration of the communication of two users [51]. The interactions among the users in a community are a significant factor for evaluating trust. The users who contribute more to the community should have a higher trust value compared to the other users who contribute less to the community. Also, two users communicating frequently in the community indicates high mutual trust between them.

In LBSNs, users can check in at POIs and we leverage the check-in activities into the recommendation. We will present the details about employing the check-in activities into the proposed recommendation approach in Section 5.3.

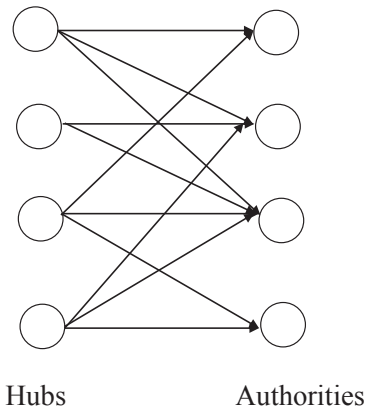

Figure 5.2: Hubs and authorities in HITS

5.2.3 HITS Model

Kleinberg proposed HITS algorithm for web search problem in [85]. HITS is also known as hubs and authorities and aims to extract information from link structures. HITS identifies good authorities and hubs for a topic by assigning two values to a page: an authority score and a hub score. These two scores are defined recursively, *i.e.*, the page's authority score is the sum of the hub values of the pages that point to it and its hub score is the sum of the authority values of the pages that it points to. Therefore, hubs and authorities exhibit a mutually reinforcing relationship and it is obvious that a good authority page is pointed by many good hubs and a good hub page points to many good authority pages, as shown in Figure 5.2. Let A be the adjacency matrix of a graph, A^t be the transpose of matrix A, a be the authority score vector and h be the hub score vector. Then the update operations of a and h are as follows [85]:

$$\begin{cases} a = A^t \cdot h \\ h = A \cdot a \end{cases} \tag{5.1}$$

In the HITS model, all the nodes are the same, *i.e.*, all are webpages. However, the reference relations among the webpages can be taken as the interactions. The concepts of

hub and authority inspire us to map the two different kinds of nodes in LBSNs to hub and authority, respectively. That is, we can consider the users as hubs and POIs as authorities. Thus, this model motivates us to employ the check-in activities in evaluating the reputation of POIs. We will describes the details in Section 5.3.

5.3 THE HYBRID TRUST-BASED POI RECOMMENDATION

5.3.1 Preliminaries

In LBSNs, the interactions between the users and POIs are important for POI recommendations. We adopt the idea of interaction-based trust model in our proposed POI recommendation. The idea behind this model is that a good POI must be a POI whose visitors' trust values are high. Also, if a user u visits many POIs with good reputation, then this user's trust value will also be high. Therefore, by exploiting the check-in records, we capture the reputation of a POI by the trust values of its visitors, and at the same time the reputation of a POI is also determine by the trust values of its visitors. This is very similar with the nature of the HITS model. We define the trust value of a user u as t_u. A simple and intuitive model for this idea can be described as follows:

$$\begin{cases} R_p = B^t \cdot T_u \\ T_u = B \cdot R_p \end{cases} \tag{5.2}$$

Here R_p is the vector of POI reputation values r_ps and T_u is the vector of user trust values t_us, B is the matrix defined based on the graph $C(U, P, E_c)$ and B^t is transpose of B.

Using this model we can calculate the POI reputation based on the interactions between users and POIs. However, the social impacts is not considered in this model. In fact, the social aspects are important in the POI recommendation. For instance, if a friend recommend a POI to a user, the user will visit the recommended POI with high probability. The reason behind that is because the user trust his friends. Thus, we can extend the equation (5.2) to incorporate the social information. Our idea is to use the trust propagation as used in the graph-based trust model into the proposed model, *i.e.*, the trust value is determined by both

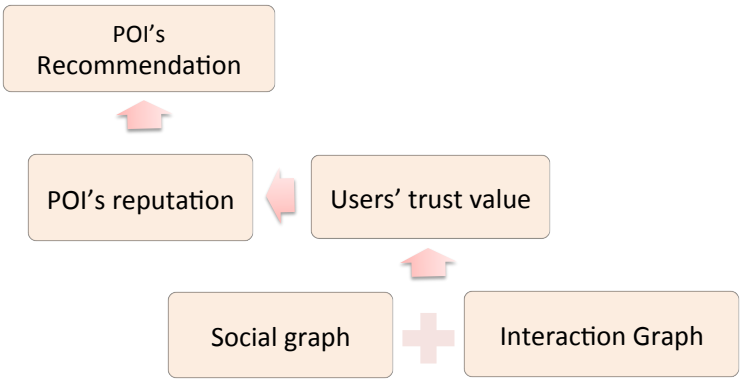

Figure 5.3: The proposed POI recommendation approach

the social graph and the reputation of the POI that the user visits. We show the model as follows; D is a matrix built based on the social graph $S(U, E_u)$.

$$\begin{cases} R_p = B^t \cdot T_u \\ T_u = D \cdot T_u + B \cdot R_p \end{cases} \tag{5.3}$$

This approach aligns with the proposed hybrid trust-based trust model, since it combines the graph-based trust model and the interaction-based trust model. The graph based trust model indicates the trust will be affected by the social network structure. Moreover, the check-in activities in LBSNs will also impact the trust of the LBSN users to POIs.

In summary, Figure 5.3 illustrates our POI recommendation approach to recommend the POIs based on their reputation. The POI's reputation value is determined by the trust values of the visitors. The trust values of the visitors are derived from two parts–the social graph and the check-in interactions created by the visitors on POIs.

5.3.2 Hybrid Trust-Based POI Recommendation

In the previous subsection, we presented our approach to combine the graph-based trust model and interaction-based trust model for a hybrid trust model for POI recommendation.

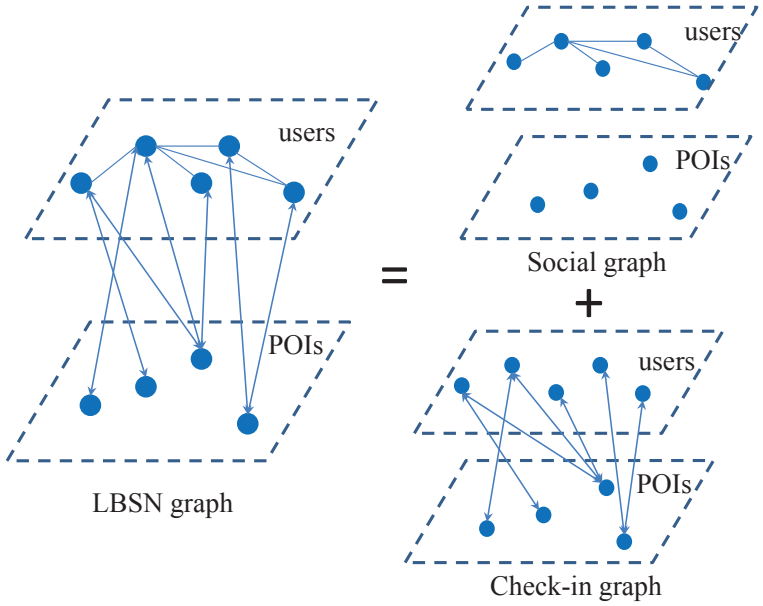

Figure 5.4: LBSN graph.

In this subsection, we present the proposed personalized hybrid trust-based POI recommendation in detail.

Before introducing our proposed hybrid trust-based POI recommendation approach, we first present the network model used in this chapter. We use a graph to represent a LBSN network. As shown in Figure 5.4, there are two different kinds of vertices in a LBSN–users and POIs. Thus, there are also two different kinds of links in a LBSN–the friendship links among users and check-in links at POIs created by users. Therefore, the LBSN graph, as shown in the left side of Figure 5.4, can be defined as $G = (V, E)$, where $V = \{v_1, v_2, ..., v_m, v_{m+1}, ..., v_{m+n}\}$ and $e_{i,j} = (v_i \leftarrow v_j) \in E$ if $v_i \in V$ is linked to $v_j \in V$ for $1 \leq i, j \leq m+n$. Here, m is the total number of the LBSN users and $U = \{v_1, v_2, \ldots, v_m\}$ are LBSN user vertex set; n is the total number of POIs in the LBSN and $P = \{v_{m+1}, \ldots, v_{m+n}\}$ are POI vertex set. Therefore the total number of vertices is $\mid V \mid = m + n$. The two types of edges in G is shown in the undirected social graph of Figure 5.4. In the social graph, the

adjacency matrix among the users can be described by

$$A_u^{m\times m}\begin{cases}1, & \text{if } e_{i,j}\in E\\ 0, & \text{otherwise}\end{cases} \tag{5.4}$$

where $1 \leq i, j \leq m$. Another type is the edges between the users and POIs, which is shown in the directed check-in graph in Figure 5.4. In the check-in graph, the adjacency matrix of the check-in graph can be described by

$$A_c^{m\times n}\begin{cases}1, & \text{if } e_{i,j}\in E\\ 0, & \text{otherwise}\end{cases} \tag{5.5}$$

where $1 \leq i \leq m$ and $m+1 \leq j \leq m+n$.

Based on the idea presented in Section 5.3.1, in the proposed hybrid trust-based POI recommendation approach,POI reputations and user trust values are computed as follows:

$$\begin{cases} r_{POIs} = (1-\beta)W_{u\text{-}POI}t_{users} \\ t_{users} = (1-\beta)W_u t_{users} + (1-\beta)W_{POI\text{-}u}r_{POIs} + \beta t^0_{users} \end{cases} \tag{5.6}$$

where $0 < \beta < 1$, $W_{u\text{-}POI}$ is the weights of the edges from users to POIs, $W_{POI\text{-}u}$ is the weights of the edges from POIs to users and W_u is the weights of friendship edges. r_{POIs} is the vector of the reputations of POIs and t_{users} is the vector of users' trust values. t^0_{users} is the initial trust value vector, and for a certain user, the initial user trust value for himself is 1 and 0 for all others. The recommendation algorithm will recommend the POIs with the top reputation values to users. Next, we describe two methods for defining the weights of the proposed approach.

5.3.3 Weight Functions

In this section, we present two different weight functions that we use in the hybrid trust-based trust POI recommendation. One is defined based on the uniform edge weights and the other is defined on non-uniform weights.

5.3.3.1 Uniform Edge Weights First, we consider the uniform distribution of the edge weights in G, which is the simplest case. For the social graph, the weights of the friendship edges are defined as:

$$W_u = \begin{cases} \frac{1-\lambda}{deg(v_i)}, & \text{if } deg(v_i) > 0 \text{ and } N_{v_i} > 0 \\ \frac{1}{deg(v_i)}, & \text{if } deg(v_i) > 0 \text{ and } N_{v_i} = 0 \\ 0, & \text{otherwise} \end{cases} \tag{5.7}$$

where $0 < \lambda < 1$, $deg(v_i)$ is the degree of user v_i in the social graph and N_{v_i} is the number of POIs that user v_i has checked in at. The weights of edges from users to POIs in the check-in graph are defined as:

$$W_{u-POI} = \begin{cases} \frac{\lambda}{N_{v_i}}, & \text{if } deg(v_i) > 0 \text{ and } N_{v_i} > 0 \\ \frac{1}{N_{v_i}}, & \text{if } deg(v_i) = 0 \text{ and } N_{v_i} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{5.8}$$

In the same way, the weights of edges from POIs to users are defined as:

$$W_{POI-u} = \frac{1}{N_{POI_j}} \tag{5.9}$$

where N_{POI_j} is the number of users who have checked in at POI_j.

5.3.3.2 Non-uniform Edge Weights For the social graph, the edge weights can be defined based on the trust level of the friendship edge. That is:

$$W_u = \begin{cases} \frac{(1-\lambda)T_{i,j}}{\sum_i T_{i,j}}, & \text{if } deg(v_j) > 0 \text{ and } N_{v_j} > 0 \\ \frac{T_{i,j}}{\sum_i T_{i,j}}, & \text{if } deg(v_j) > 0 \text{ and} N_{v_j} = 0 \\ 0, & \text{otherwise} \end{cases} \tag{5.10}$$

where $0 < \lambda < 1$, $T_{i,j}$ is the trust value indicating how much user j trusts user i. In our dataset, such trust values between users are not available, so we still use the uniform friendship edge weights in the experiment.

Regarding the weights of edges between the POIs and users, we propose two methods for determining these weights. The first method employs the check-in counts to define the weights and the second method employs the entropies to define the weights.

- *Weights based on check-in counts*: The weights of edges from users to POIs in the check-in graph can be defined based on the number of check-ins at the POI created by the users:

$$W_{u-POI} = \begin{cases} \frac{\lambda C_{i,j}}{\sum_i C_{i,j}}, & \text{if } deg(v_j) > 0 \text{ and } N_{v_j} > 0 \\ \frac{C_{i,j}}{\sum_i C_{i,j}}, & \text{if } deg(v_j) = 0 \text{ and } N_{v_j} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{5.11}$$

  where $C_{i,j}$ is the number of check-ins created at POI i by user j. In the same way, the weights of edges from POIs to users are defined as:

$$W_{POI-u} = \frac{C_{i,j}}{\sum_j C_{i,j}} \tag{5.12}$$

- *Weights based on entropies*: We define the user set as U, and the set of the POIs in the LBSN as P. We define $P_{i,j}$ as the probability that user j checked in at POI i given all of j's check-ins; that is, it is the fraction of the number of check-ins of user j created at POI i over the total number of check-ins of user j. So the user entropy of user j can be defined as in Equation 5.13. We define $P'_{i,j}$ as the probability that the user j checked in at POI i given all the check-ins at i; *i.e.*, it is the fraction of the number of check-ins of

user j created at POI i over the total number of check-ins at POI i. So the POI entropy of POI i can be defined as in Equation 5.14.

$$E_{u_j} = -\sum_{POI_i \in P} (P_{i,j}) \times \log P_{i,j} \tag{5.13}$$

$$E_{POI_i} = -\sum_{u_j \in U} (P'_{i,j}) \times \log P'_{i,j} \tag{5.14}$$

Thus, the weights of edges from users to POIs in the check-in graph can be also defined based on POI entropies:

$$W_{u-POI} = \begin{cases} \dfrac{\lambda P_{i,j} \log(P_{i,j})}{E_u}, & \text{if } deg(v_j) > 0 \text{ and } N_{v_j} > 0 \\ \dfrac{P_{i,j} \log(P_{i,j})}{E_u}, & \text{if } deg(v_j) = 0 \text{ and } N_{v_j} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{5.15}$$

In the same way, the weights of edges from POIs to users are defined as:

$$W_{POI-u} = \frac{P'_{i,j} \log(P'_{i,j})}{E_{POI}} \tag{5.16}$$

5.4 DATA ANALYSIS

The dataset used in this chapter is not the same as that in Chapter 3 and Chapter 4. First, the check-in dataset used here are collected from March, 2012 to July, 2012. During this period, we gather 1,226,769 check-ins created by 44,437 Foursquare users at 17,816 POIs. Thus, the average check-ins per user (AC_u) is 27.61 and the average check-ins per POI (AC_p) is 68.86. Moreover, the social network of the users are used in this chapter and we obtain 297,580 friendship links among the 44,437 users. We present the summarized dataset in Table 5.1. We also summarize the POIs on the aspect of top 9 categories[1] in Table 5.2.

Next, we investigate the complementary cumulative distribution function (CCDF) of the check-ins created at the POIs of this dataset and show the result in Figure 5.5. From

[1]There are POIs whose category property is null.

Table 5.1: Summary of check-in and social network dataset

Number of POIs	17,816
Number of Users	44,437
Number of Check-ins	1,226,769
Friendship Links	297,580
Average Friends per User	6.7
AC_u	27.61
AC_p	68.86

Table 5.2: Summary of POIs on the aspect of top 9 categories of dataset in Chapter 5

Category	# of POIs	# of Check-ins	AC_u	AC_p
Arts & Entertainment	462	60,685	3.62	131.35
College & University	753	62,740	11.31	83.32
Food	2,359	250,540	19.78	106.21
Professional & Other Places	2,940	156,239	9.34	53.14
Nightlife Spot	877	106,338	6.36	121.25
Great Outdoors	861	57,674	5.18	66.98
Shop & Service	3,273	227,382	12.25	69.47
Travel & Transport	1,233	78,709	4.01	63.84
Residence	2,846	147,076	36	51.68

Figure 5.5 we can see that about 80% of POIs obtain at least 5 check-ins during our data collection period. We also can observe that there are about 15% of POIs which have more than 100 check-ins and these are popular POIs. Moreover, we find that there are around 0.1% of POIs with more than 2,000 check-ins. These numbers indicate that popular POIs attract a lot of check-ins. Our overall venue popularity analysis in Section 3.2.1 based on cumulative dataset also imply the similar result, thus in LBSNs, popular POIs seems to be an important factor in attracting LBSN users. Therefore, we should consider this factor into our proposed personalized POI recommendation. We also explore the CCDF of the check-ins created by users. As shown in Figure 5.6, we can see that there is only about 30% of users who have created more than 10 check-ins and there is only around 7% of users who have created more than 100 check-ins. Such a result indicates that users do not frequently share their check-ins in Foursquare. The social graph in our dataset contains 44,437 users and 297,580 friendship links among them, so we also study the properties of such social relations and plot the CCDF of users' friendship links in Figure 5.7. The average friend links in our dataset is 6.7 and the maximum friend links is 773. Figure 5.7 shows that there is only around 20% of users who have at least 10 friends.

We introduce the concepts of user entropy and POI entropy in Section 5.3.3.2 and we study the user entropy based on the number of check-ins created by the user, as well as the POI entropy based on the number of check-ins at the POI.

The average user entropy and average POI entropy based on the check-in counts are shown in Figure 5.8 and Figure 5.9, respectively. Usually the concept of entropy is used to measure the randomness of the activities [86]. Therefore, we employ the user entropy to measure the diversity of the users' check-in activities and use the POI entropy to measure the diversity of the users who have checked in at a POI [87]. In Figure 5.8, we can see that for users who have fewer than 100 check-ins, their user entropies keep rising with the increase in the number of check-ins. For the users who have more than 100 check-ins, the user entropies of some of them still keep rising, which indicates these users like exploring new POIs. However, there are still some users whose user entropies do not increase with the growth of the number of check-ins, which indicates that they prefer to check in at the POIs they have been to before. We note that only around 7% users in the dataset have more than

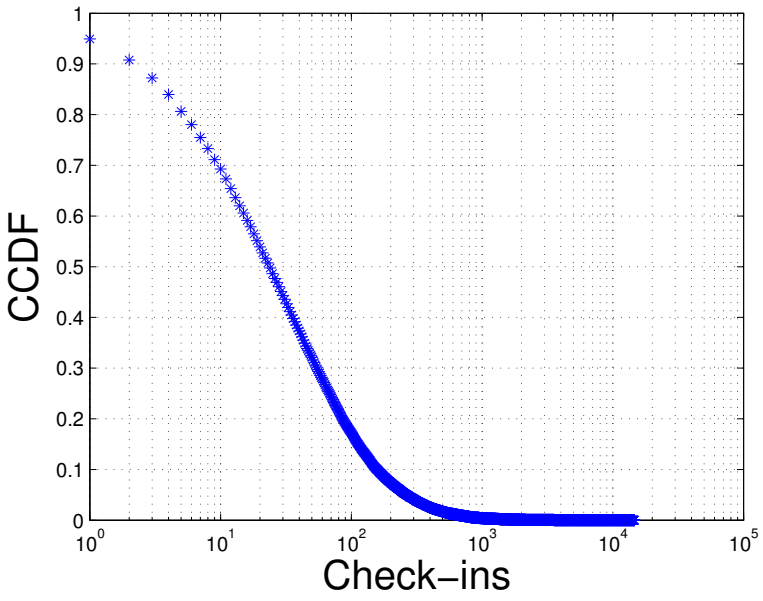

Figure 5.5: CCDF plot of the check-ins created at the POIs of check-in dataset in Chapter 5

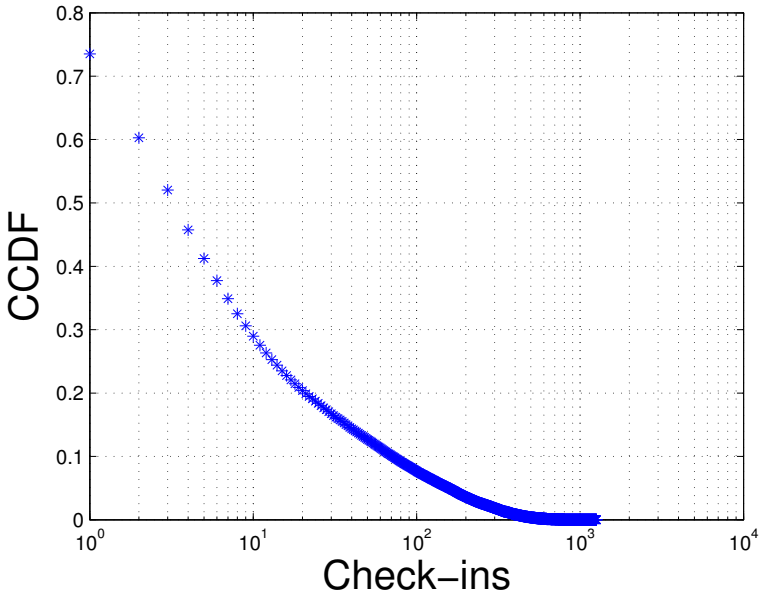

Figure 5.6: CCDF plot of the check-ins created by users of check-in dataset in Chapter 5

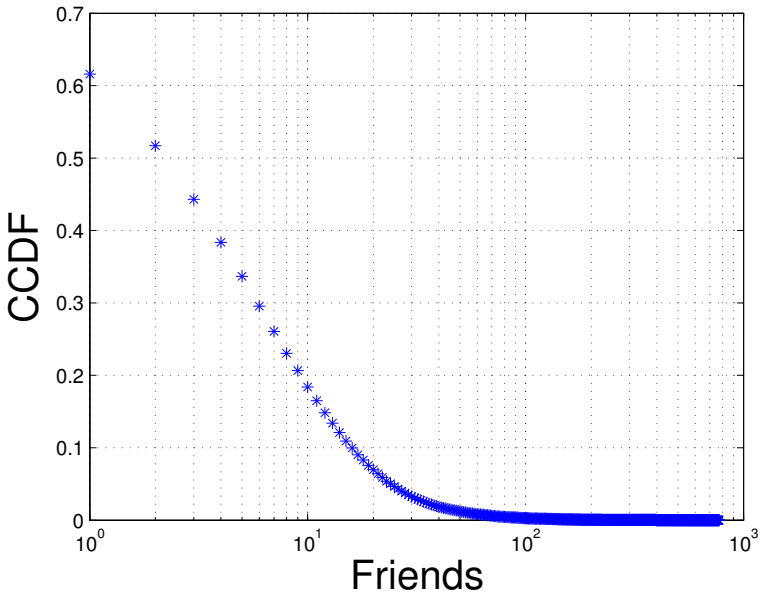

Figure 5.7: CCDF plot of the Friendship of social network dataset

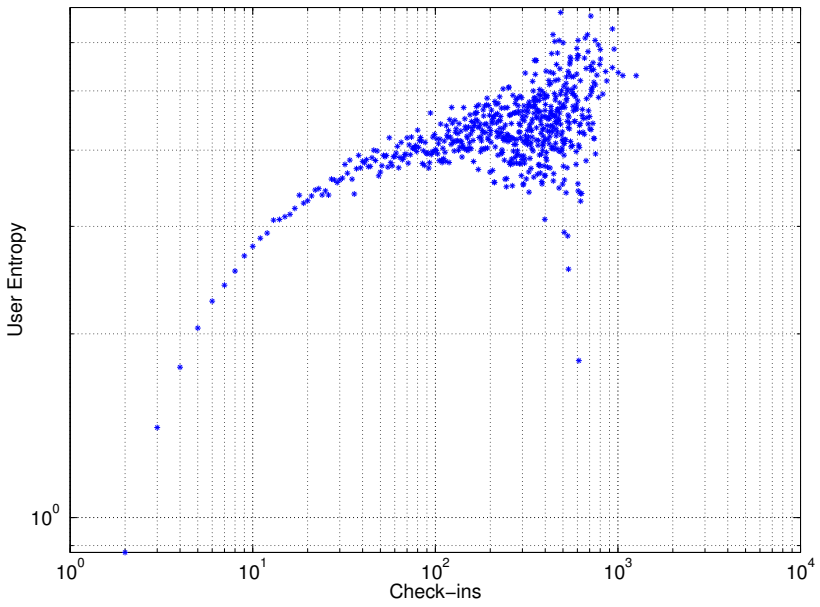

Figure 5.8: Average user entropy based on the check-in counts

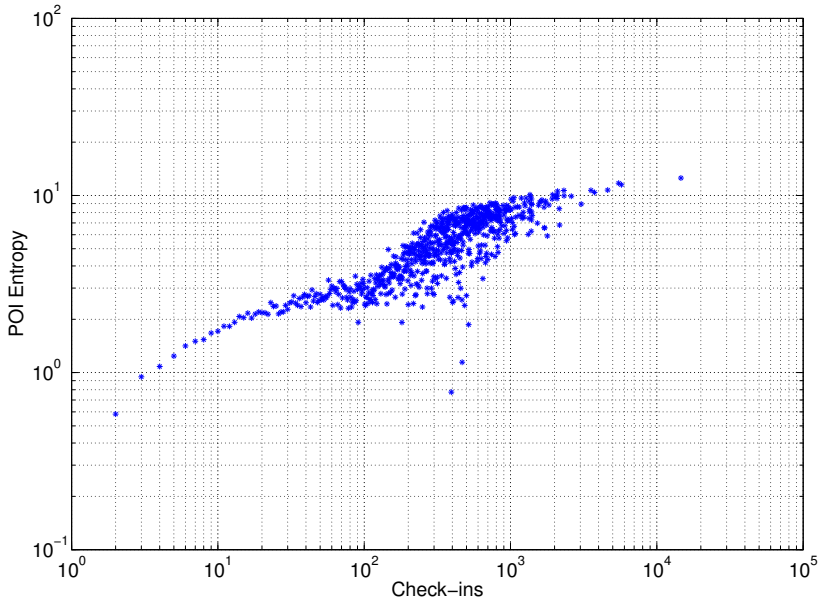

Figure 5.9: Average POI entropy based on the check-in counts

100 check-ins, therefore, the majority of users exhibit diversity in their check-in activities. From Figure 5.9 we can see that generally the POI entropy increases with the growth of the number of check-ins created at the POI. This trend is especially notable for the POIs with no more than 50 check-ins. According to Figure 5.5, around 70% of the POIs have no more than 50 check-ins, therefore the POI entropies of most of the POIs exhibit an increasing trend, which implies that POIs usually attract different users. Therefore, the data analysis about the POI entropy tells us that in LBSNs, popular POIs always attract more users; the data analysis about the user entropy implies that users tend to visit new POIs. Thus, in POI recommendation it is very important to recommend new POIs to users.

5.5 EXPERIMENTS

In this section, we evaluate the proposed hybrid trust-based POI recommendation approach on our Foursquare dataset presented in Section 5.4. We first introduce the methodology

and measures that we use to evaluate the proposed hybrid trust-based POI recommendation approach, and then we present our results.

5.5.1 Methodology and Measures

Table 5.1 summarizes our dataset that has more than one million check-ins but is very sparse as many users checked in at a POI more than once. Thus, we first pre-process the dataset by removing the users and POIs with very few check-ins, and then randomly divide the remaining dataset into a training dataset (80%) and a testing dataset (20%) to evaluate the approach. In particular, we run the proposed approach on the training dataset, and use the reputation value of the POIs to rank the POIs. Since the entropy related analysis shows that users like to visit new POIs, the POI recommendation in this section only recommend the new POIs. That is, we recommend the POIs that the users haven't visited before in this section. We make the *top-N* POI recommendations based on the ranks of the POIs. Then we compare the recommended *top-N* POIs with the testing dataset. We use the precision ($P@N$) and recall ($R@N$) measures to evaluate the approach, which are defined as follows:

$$P@N = \frac{|\text{relevant documents} \cap \text{retrived documents}|}{|\text{retrived documents}|} \tag{5.17}$$

$$R@N = \frac{|\text{relevant documents} \cap \text{retrived documents}|}{|\text{relevant documents}|} \tag{5.18}$$

We first evaluate the $P@N$ and $R@N$ for the proposed POI recommendation on the entire dataset and compare results with the random walk with restart (rwr) recommendation approach proposed in [11]. The random walk with restart can provide a personalized view of the network and can be adopted to personalized rankings [88]. As presented in [11], the transition probabilities can be obtained by $Q = \alpha W + (1 - \alpha)R$. Here W is determined based on the network structure, R presents a random probability of jumping to any other node and α is a parameter to tune the behavior [11]. Then, the steady-state probability vector p can be calculated by $p = pQ$ and the recommendation is made based on p. We choose to compare with rwr because rwr approach performs best as shown in [11] and we

use the recommended parameter in [11] for *rwr*[2]. Then, we also evaluate the two measures of the proposed approach and *rwr* approach on the categorized dataset. Lastly, we investigate the performance of the proposed approaches based on different parameters β and λ. In the experiment, for the user i, the initial hub score vector is defined as a vector with all zeros except the i^{th} element, which is 1.

5.5.2 Results

5.5.2.1 Results using the entire dataset In this subsection, we use the entire dataset, *i.e.*, we do not consider the categories of POIs in the dataset. Here we choose $\beta = 0.85$ and $\lambda = 0.95$ (we will study these parameters in Section 5.5.2.3). In Figure 5.10, we can see the average *P@N* of the hybrid trust-based POI recommendation approach using entropy weights is the highest, followed by that of the hybrid trust-based POI recommendation approach using the uniform weights. In Figure 5.11, we can also see that the *R@N* of the hybrid trust-based POI recommendation approach using entropy weights is the best, followed by that of the hybrid trust-based POI recommendation approach using the uniform weights. All the three proposed approaches perform better than *rwr* with regards to both *P@N* and *R@N*. For all the four approaches, the performance of *R@N* gets better with increasing values of N.

5.5.2.2 Results using categorized dataset In this subsection, we evaluate the *P@N* and *R@N* in Nightlife Spot category dataset by using the same parameters as used in Section 5.5.2.1. The average *P@N* results are shown in Figure 5.12. We can see that the average *P@N* of the proposed approach using entropy weights is still the highest compared to that of the others, followed by that of the proposed approach using the edge weights based on the check-ins. As shown in Figure 5.13 for *P@N*, we can also see that the proposed approach using entropy weights is the highest compared with the others, followed by the proposed approaches using the edge weights based on the check-ins. Moreover, the *R@N* increases with increasing values of N and there is little difference between the results for the pro-

[2]Changing the parameter of *rwr* will get better results, however, our proposed approach using entropy weights is still better.

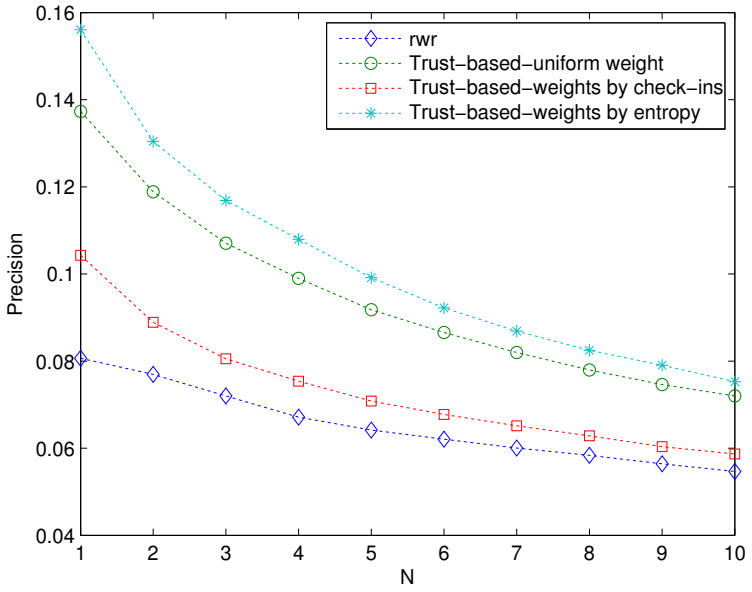

Figure 5.10: Average $P@N$ on the entire dataset

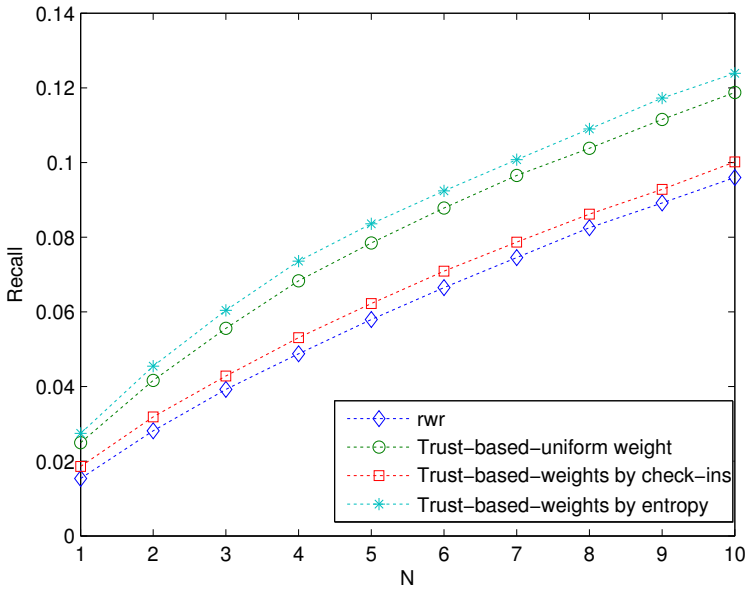

Figure 5.11: Average $R@N$ on the entire dataset

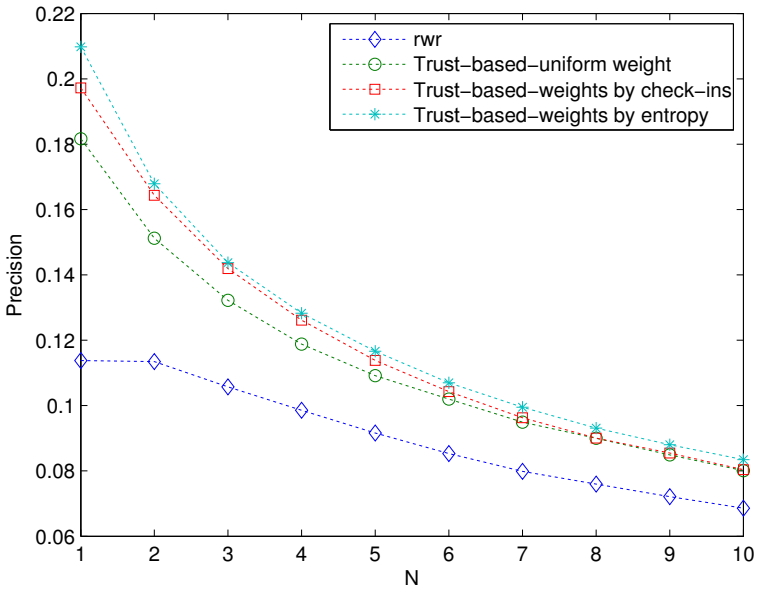

Figure 5.12: Average $P@N$ on the Nightlife Spot dataset

posed approaches using different edge weights. Both of them achieve better results than the proposed approach using uniform weights. Note that when using the entire dataset, the proposed approach using uniform weights performs better than that using weights based on check-ins. The reason for such a difference may be that in the entire dataset, users' check-ins are more diverse than in the single categorized POIs and the category information brings more unpredictable factors to the recommendation. Therefore, the uniform distribution of the weights may achieve better results. However, regarding the dataset in a single POI category, the number of the check-ins reflects the reputation of the POIs more accurately as we are comparing the POIs in the same category. Thus, the approach with the weighted edge will give better results than using the uniform weight. An example to explain the difference is to consider an office building and a restaurant, as more check-ins in the building POI does not indicate it is a better recommendation than the restaurant. Thus, it is not surprising to see that the weighted edge approach performs better than the uniform edge approach if we only use the categorized dataset.

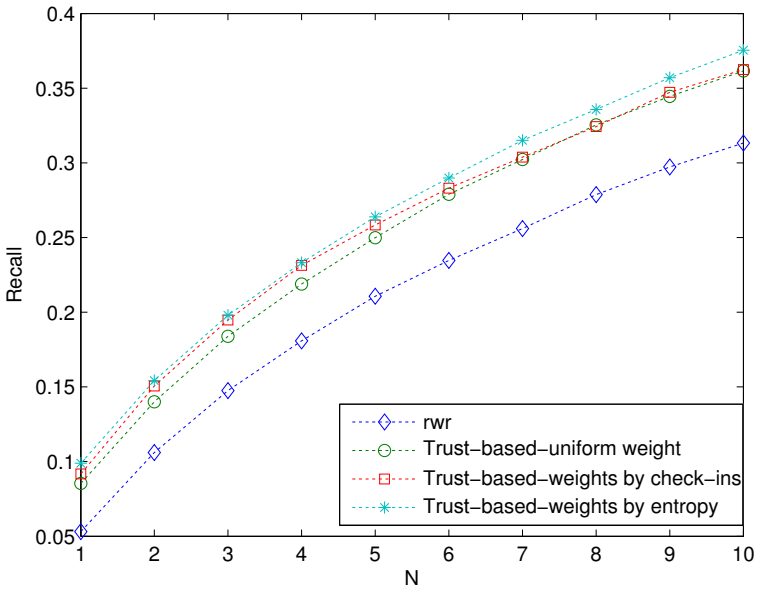

Figure 5.13: Average $R@N$ on the Nightlife Spot dataset

5.5.2.3 Parameter Setting In the proposed hybrid trust-based POI recommendation approach, there are two parameters β and λ. Generally, with different datasets, β and λ should be set according to the properties of the dataset to get the best results. Here we discuss the impacts of different values of parameters β and λ. We show results of $P@N$ using the proposed approach with weights based on entropy by changing β and λ on Nightlife Spot category dataset in Figure 5.14. The best performance based on $P@N$ is achieved by $\beta = 0.85$ and $\lambda = 0.95$; $\beta = 0.15$ and $\lambda = 0.95$ give the worst performance. Generally, we can see that larger β performs better than smaller β. The parameter β controls the importance of the initial hub score vector. In our experiment, for a user i, the initial hub score vector is all zeros but the i^{th} element is 1, which indicates that the nodes which are closer to user i contribute more to the hub score update. However, setting parameter λ is more complicated than β. The parameter λ balances the impacts of social influence and the check-in behavior in updating hub scores. The larger λ indicates that the check-in behavior contributes more in hub score updates. When β is very small (*e.g.*, $\beta = 0.15$), larger λ gives worse results; however, when β is very large (e.g. $\beta = 0.85$), larger λ gives better results. In summary, the

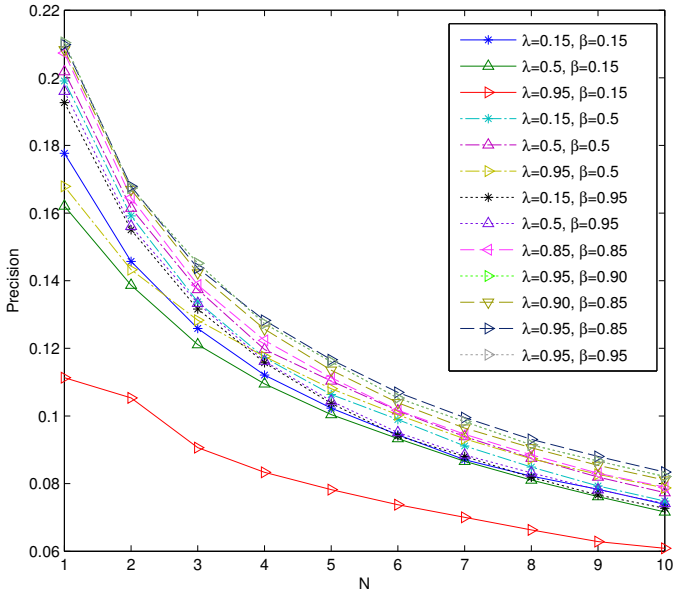

Figure 5.14: Average $P@N$ on the Nightlife Spot dataset by different parameters.

optimal parameters of β and λ should be set based on the properties of the dataset.

5.6 DISCUSSION

In this chapter, we have proposed a hybrid trust model for POI recommendation in LBSNs. The goal of the POI recommendation is to give the personalized POI recommendation to individual users in LBSNs. We would like to recommend POIs with good reputation to the LBSNs users thus we propose a trust-based POI recommendation approach to compute the POI reputation for each individual users. The POI reputation is determined by the LBSN users' check-in behaviors, which can be considered as the interaction-based trust model. Since our goal is a personalized recommendation, thus, we consider the social relationship (graph-based trust model) in evaluation of the trust values of the LBSN users. The hybrid trust model is one contribution of the proposed approach.

Another contribution of this proposed approach is that it proposed a generalized ap-

proach to evaluate the reputation of nodes other than user nodes. In the graph-based trust model, all the nodes are user nodes and the trust evaluation are based on all homogeneous type of nodes (*i.e.*, user nodes). Although the interaction-based trust model employ the interactions between two different types of nodes (*e.g.*, users nodes and email nodes), such models still focus on trust evaluation of the user nodes. Obviously, evaluating and calculating trust of user nodes in social networks or other B2C platforms are important, as it can be used in detecting and preventing various frauds or attacks. Moreover, evaluating the reputation of object nodes other than user nodes can benefit many applications, especially in recommendation systems. There is few work on evaluating the reputation of object nodes by employing interactions between the object nodes and user nodes, as well as the social relationship among the user nodes. Thus, the proposed work contributes in this area. Although the proposed work aims at personalized POI recommendation, it can be extended to many similar platforms or applications. For example, personalized item recommendation in online shopping website. The purchase history can be taken as the interactions between the users and items on the platform. With the social information among the users (if it is available), the proposed model can be extended for personalized item recommendation.

Moreover, the proposed hybrid trust-based POI recommendation is also helpful for event/location planning for a group of users. Assume a scenario that several friends want to find a POI that all/most of them would be interested in to hang out together, the proposed approach can be easily used in this case. By simply changing the initial user trust value vector and then running the proposed approaches would get the ranked list of POIs. That is, for recommending POIs to a set of users $\{v_{x_1}, \ldots, v_{x_n}\}$, the initial user trust value vector should be $t_i^0 = \begin{cases} \frac{1}{n} & \text{if } i \in \{x_1, \ldots, x_n\} \\ 0 & \text{otherwise} \end{cases}$. In this case, each user in the group has the same level of preference in finding the common POI. If the scenario is that a group of people would like to celebrate a birthday or other special event for a certain user or some users, these certain users' preferences should have higher priority than others. Our proposed model can be also used in this case, by simply adjusting the initial users trust values from the same value (*e.g.*, $\frac{1}{n}$) to different values. That is, assign a higher number to the users whose opinion are more important in making the decision.

Therefore, the proposed hybrid trust-based model does not only cover the shortage of trust model area, but also provides a generalized approach in evaluating the trust value of heterogeneous nodes through social information and the interactions between the two types of heterogeneous nodes. The experimental results show good performance of the proposed approach in LBSNs field, however, it can be extend to other fields. Moreover, the personalized recommendation is not limited in recommending POIs for an individual user, but provide a possible solution and guidance to recommendations for a group of users.

5.7 SUMMARY

In this chapter, we first present the motivation of our proposed hybrid trust-based POI recommendation. In particular, we describe how the graph-based trust model, interaction-based trust model and HITS model inspire our proposed work in POI recommendations in LBSNs. The HITS model provides a wonderful approach in evaluating the rank of the webpages through the references among the hyperlinks. However, it cannot be directly used in social platforms as it does not consider the social information. So we propose combining the graph-based trust model and the interaction-based model for a hybrid trust-based model. Then, we formalize the POI recommendation in LBSNs and explain our proposed hybrid trust-based POI recommendation model in details. We also present several weight function definitions related to POI recommendation scenarios in LBSN platform. Then, we analyze our dataset and find that users trend to visit new POIs thus in our experiment we focus on the performance of the proposed model in new POI recommendations in LBSNs. In Section 5.5, we narrate our methodology and measures to evaluate the performance of the proposed model on Foursquare dataset. We present the detailed experimental results of the proposed model and compare it with another POI recommendation approach. The experiments show good performance of our proposed model and better results than the other approach. We also investigate the parameter setting in this section. Then, we discuss the contributions of the proposed hybrid trust-based model. In particular, it provides a generalized approach to evaluating the trust value in a network with heterogeneous nodes.

6.0 CONCLUSIONS AND DISCUSSIONS

Recently, social media and mobile technology have become increasingly important parts of our society. LBSNs, that combine mobile technology, location-based services and social media, have been changing the way that users use location information in social life. Thus, investigating LBSN dataset, especially the dataset of a specific city (due to the physical geographic limitation of human activities), is extremely helpful and valuable to:

- Understand human preferences in POIs and then design better marketing strategies for POIs.
- Understand the human mobility patterns to achieve better urban designs and plans.
- Understand the factors that impact human activities related to physical locations, in order to design better POI recommendation approaches.

In this dissertation, we have carried out various studies towards fulfilling these objectives has been done in this dissertation. That is, based on the real dataset collected from a popular LBSN (Foursquare), we first analyze the entire online population's location preferences based on the cumulative dataset. Then, we mine the dataset through a probabilistic graphical model (LDA) to investigate the human mobility patterns and lastly we propose a hybrid trust-based model to recommend POIs for individual users.

In this chapter, we summarize the contributions of the proposed work and discuss how they address the challenges proposed in Section 1.1 and the limitations. Moreover, we enumerate some future research directions based on the work proposed in the dissertation.

6.1 CONTRIBUTIONS

In this dissertation we focus on LBSN users' behavior to investigate the users' preferences based on real LBSN data. Our study first examines the entire online population's preferences of locations based on the cumulative number of check-in information, *etc.*, from the Foursquare dataset pertaining to the greater Pittsburgh area. Then, we focus on the users' check-in records created on POIs and mine such dataset through LDA model to obtain the geographic topics of the POIs based on users' mobility behavior. The acquired geographic topics indicate part of online users' geographic preferences as a cluster of locations. Lastly, we work on personalized POI recommendations for LBSN users based on their historical check-in behavior, as well as the social relationship with their friends. The aforementioned three parts of work in this dissertation successfully address the challenges proposed in Section 1.1. Therefore, we will present below the contributions of the three parts of work and how they address the challenges mentioned before.

- Challenge I: understanding local venue popularity in LBSNs from the view point of the entire online population. We have done a comprehensive and quantitative analysis of the venue popularity to address this challenge in Chapter 3. In particular, we study the overall venue popularity and overall category popularity based on the cumulative dataset collected from Foursquare. As far as we know, this is the first work to investigate the local venue popularity in LBSN area based on the cumulative dataset. Current work in this area usually employs the check-in data collected during a certain period, however, the lack of the information before the data collection period could not provide a complete overview of the overall venue popularity. Although our cumulative data is just about the number of check-ins and the number of users visited at the venues, it shows the overall venue popularity not just within a period. Therefore, the proposed analysis provides a more complete understanding about venue popularity. Moreover, we explore the spatial and temporal features of the trending venues and study the impacts of Foursquare features such as specials, web presence and menu on promoting the venue popularity. The results are very helpful for designing marketing strategy for the venue owners in LBSNs. Our work presents a quantitative analysis about the local venue popularity in LBSNs, which

is based on the entire online population's check-in information. This analysis provides an overview of the entire online population's preferences on locations. Since the locations are closely related to human activities, it is very valuable and helpful in understanding human mobility behavior.

- Challenge II: exploring latent geographic topics based on the movements of unobserved groups of users. Through the LDA model, we explore the user mobility-driven geographic topics in this dissertation. The geographic topics exhibit a certain group of users' preferences of locations. The topic results (*e.g.*, the sports topic, two topics related to university and so on) shown in our experimental results reveal clusters of locations based on some unobserved groups of users' trajectories (*e.g.*, university students and sports fans). The latent geographic topics successfully answers the questions asked in Section 1.1.2. For example, the cafe found in the topic related to PITT is the popular one for the PITT freshman who want to find the popular cafe among PITT community. The restaurants in the topic related to shopping and entertainment show the popular dinning places after shopping. Therefore, the latent geographic topics reveal the preferences of the unobserved group of people on locations and the mined geographic topics are capable of advancing venue recommendation and friend recommendation. In particular, we can recommend users whose latent topics are very similar as friends because of the similar interests of activities related to location or recommend the top but unvisited venues of a user's latent topic. Moreover, we explore the geographic topics formed on two different temporal aspects (*e.g.*, weekdays and weekends), and the topics especially the differences of the topics exhibit the different human mobility behavior between weekdays and weekends. Furthermore, we study the travelers' mobility behavior and this, to our best knowledge, is the first such research related to mining travelers' behavior based on LBSNs. We investigate the popular categories based on travelers' check-ins, as well as the temporal features of the check-ins as the distribution of the top 9 categories, which is helpful in understanding travelers' goals and then improve the services for travelers. Also, the latent geographic topics based on the travelers' check-ins can better help to understand the travelers' requirements and then improve the location recommendations to travelers. Thus, our work by employing LDA to mine the latent geographic topics successfully

address the challenge of understanding groups of people's preferences of locations. Compared to the work in Chapter 3, our work of mining the latent geographic topics in Chapter 4 concentrates on investigating the mobility pattern of groups of users based on their check-in records, thus exhibits deeper and further research in understanding the user behavior in LBSNs.

- Challenge III: personalized hybrid trust-based POI recommendation. After exploring the entire online population's preferences of locations and the groups of users' geographic topics, we proposed a hybrid trust-based POI recommendation to address the POI recommendation challenge. Our proposed personalized recommendation approach does not only consider the individual's check-in behavior, but also integrates the social information. Therefore, the performance of the proposed hybrid trust-based POI recommendation is very good as it captures both the social impact and user behavior impact. We also present several weight functions used in the proposed hybrid model, and such weight functions can be extended based on the specific applications. Thus, the proposed model is very flexible and can be adapted to other applications. Besides, the hybrid trust-based POI recommendation combines the graph-based trust model and interaction-based trust model, advances the limited research in hybrid trust-based model. Motivated by the existing HITS model in web search area, our proposed model effectively integrates the social impacts (graph-based model) and user historical behavior (interaction-based model) in POI recommendation. Since the hybrid trust-based model has both the advantages in graph-based model and interaction-based model and there is very limited work in this area, our work presents first exploration in this area. Furthermore, the proposed model provides a generalized approach to evaluate the reputation of object nodes other than user nodes in a network consisting of two types of nodes through the interactions between the user nodes and object nodes.

In summary, in this dissertation we have done a comprehensive work towards understanding the users' preferences of location in LBSNs, from the entire online population's perspective, groups of users' viewpoint and finally from an individual's angle. The work of this dissertation advances in exploring the user mobility behavior through the social media.

6.2 LIMITATIONS

There are some limitations of the research conducted in this dissertation. We discuss them in this section.

The online dataset may be biased by the demography information of the online population (*e.g.*, gender and age), which is a limit of using the online dataset to analyze the human preferences of locations. Moreover, we simply take the check-in as a positive sign to the location, which may not always true. The data analysis and mining about the human preferences on locations are based on the dataset collected in the greater Pittsburgh area in Foursquare, thus the scope of our work is limited by the size of the available dataset. The analysis of the online population's preferences of locations are limited to the greater Pittsburgh area and the user mobility patterns are also based on the activities of people in this area. Although the dataset used in Chapter 3 is cumulative data (*e.g.*, cumulative number of check-ins at a POI and *etc.*), which gives us an overview of the overall popularity of the POIs; we still miss the temporal features of such data. If the check-in data are available (we mean the check-ins from the POI being created), we could explore more features of the online population's preferences of locations.

Besides the limitation of the local dataset, another limitation is the data collection period. The dataset is collected from February to July, 2012, hence we miss the user activities at the beginning of 2012 and after July. Therefore, some important user mobility patterns are not observed in our study. For example, as we mentioned in Section 4.3.2.1 we do not see Heinz Field (American Football Stadium) as a top POI in Sports Topic because the data collection period does not cover the football season. Another limitation of the proposed work is about the definition of the travelers. In Section 4.4.1 we simply define a traveler as a user whose hometown is more than about 310 miles away from downtown Pittsburgh. We believe the definition of a traveler should consider more information about the user. However, we can not get more information from the public data provided by Foursquare API, thus we could not find a better way to differentiate travelers from local users. If more information about the users are available to help categorize travelers, we will achieve a more comprehensive and accurate study about the travelers' behaviors.

With regards to the proposed hybrid trust-based POI recommendation, our assumption is that all the check-ins are honest. We do not consider any threat model in evaluating users' trust values and and POIs' reputations. However, some attacks could subvert the proposed approach and sybil attack is a case in point. That is, if a malicious user creates a lot of fake accounts, cleverly creates friendship relations with other genuine accounts and also creates a large number of fake check-ins, the POIs' reputation can be impacted by such fake check-ins from the fake accounts. Thus, more work is needed in preventing such attacks. Moreover, another limitation of the proposed hybrid trust-based POI recommendation is that its convergence is not mathematically proven. Our implementation about stopping the reputation calculation is based on setting maximum iteration times and a small constant. The calculation will be stopped if achieving the maximum rounds of iteration or the difference between the current reputation value and the last value is smaller than the constant. Besides, one may doubt that the recall and precision are not as high as the item recommendations in other platforms such as Amazon. This is mainly caused by the sparse matrix of users' check-ins. Table 5.1 shows that the average check-ins per user is about 27 within 5 months, which indicates that a user engages in about one check-in per week. Thus, we can find that on average LBSN users usually share very few check-ins, which is different from the dataset collected from Amazon or Netflix. In the platform such as Amazon or Netflix, all the purchase histories are recorded. However, in LBSN platform, only the user activities shared by the users will be recorded and the most activities related to POIs are not shared by the users on LBSN platform. Therefore, the sparse dataset limits the performance of the proposed algorithm.

6.3 FUTURE WORK

In this section, we briefly enumerate the potential future research directions in the context of this dissertation.

We mentioned in the previous section that our study are based on a local dataset pertaining to the Greater Pittsburgh area. Thus, employing our data analysis and mining

approaches on other local datasets or a global dataset will strengthen the work on exploring human preferences of locations. Also, since some periods within a year are not covered by our dataset, with more data especially collected during the missing period will give a more comprehensive understanding of the human mobility patterns and more meaningful geographic topics. Moreover, if more demographic information such as gender or age information is available, employing the proposed approach to analyze and mining dateset on a basis of such information is also very interesting.

More explorations of using the SVD-based approach or other models on the LBSN check-in dataset are also interesting. In Section 4.6 we study our check-in dataset using the SVD-based approach and we get some interesting results. However, our exploration about it is still at an early stage. For example, we speculate that the sparse feature of our check-in dataset and the many 1s in the matrix may cause the SVD-based approach to capture the user behaviors or patterns related with the significant elements in the matrix, so it would be interesting to investigate the results of using SVD-based approached on normalized matrix.

Moreover, the proposed POI recommendation approach proposed in this dissertation can be extended to predict the LBSN users' next check-ins. By carefully considering the temporal aspects of the users' check-ins, a novel approach to predict the human mobility based on the social impacts and users' previous behavior can be developed. Also, temporal features could be used to extend the proposed POI recommendation approach to improve the performance of the recommendation. For example, recommending a restaurant at the lunch/dinner time is obviously better than recommending a bookstore at that time. In this dissertation we do not consider the temporal feature in the proposed POI recommendation, which could be an interesting direction in improving the approach. Furthermore, our evaluation result of the proposed hybrid trust-based POI recommendation using the categorized dataset achieve higher performance in both precision and recall. Another direction of improving the proposed recommendation approach could be considering the category information of the POIs.

In Section 5.6 we mentioned how to simply employ the proposed approach in POI recommendations for a group of users. With regards to the POI recommendations, we believe recommending POIs for a group of users is also important and will be the next hot research area. This is because the traditional item recommendation requirement usually targets in-

dividual users, *e.g.*, recommending books on Amazon or movies on Netflix; there is limited social activities involved in such requirements. However, social activities play a very important role in human activities related to locations, therefore the POI recommendations which are closely related to the human activity should be able to capture the social factors and consider the location preferences of important users attending the activity. Although the proposed hybrid trust-based POI recommendation leverage the social graph in evaluating the reputation of POIs in making recommendations, it still focuses on the POI recommendations for individual users rather than for a group of users. Recommending POIs for a group of users will be a very practical and interesting topic and our current work simply shows a possible direction toward it.

Last but not the least, the proposed hybrid trust-based model may be extended to detect the fake ratings. In some platforms, users can leave a rating of the items and it is possible for some malicious users to leave fake and high ratings for some bad items. Using the idea of the proposed hybrid trust-based model to compute the reputation of the items may be helpful in detecting the suspicious items, as the reputation calculated based on the honest interactions and social information (if available) will be different with the fake high ratings. Therefore, it would be interesting to explore using the real and honest interactions and social information to address such fake rating issues.

BIBLIOGRAPHY

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[2] Facebook. [Online]. Available: http://touch.facebook.com/

[3] Foursquare. [Online]. Available: https://foursquare.com/

[4] (2012) 200m users include location in facebook posts; company looks to expand location apis. [Online]. Available: http://www.insidefacebook.com/2012/04/05/200m-users-include-location-in-facebook-posts-company-looks-to-expand-location-apis/

[5] (2015) About us. [Online]. Available: https://foursquare.com/about/

[6] The full list of foursquare badges. [Online]. Available: http://www.4squarebadges.com/foursquare-badge-list/

[7] (2011) Foursquare badges level up to encourage exploration. [Online]. Available: http://sproutsocial.com/insights/foursquare-badge-update/

[8] (2011) How important is foursquare to your business? [Online]. Available: http://conversationalmarketinglabs.com/blog/2011/11/social-media-2/how-important-is-foursquare-to-your-business/

[9] (2011) The importance of foursquare for a business. [Online]. Available: http://www.htmlgraphic.com/the-importance-of-foursquare-for-a-business/

[10] B. Berjani and T. Strufe, "A recommendation system for spots in location-based online social networks," in *Proceedings of the 4th Workshop on Social Network Systems*, ser. SNS '11. New York, NY, USA: ACM, 2011, pp. 4:1–4:6.

[11] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo, "A random walk around the city: New venue recommendation in location-based social networks," in *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, ser. SOCIALCOM-PASSAT '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 144–153.

[12] What is a speical? [Online]. Available: http://support.foursquare.com/entries/195165-What-is-a-special-

[13] C. Song, T. Koren, P. Wang, and A.-L. Barabasi, "Modelling the scaling properties of human mobility," *Nat Phys*, vol. 6, no. 10, pp. 818–823, 10 2010.

[14] F. Simini, M. C. Gonzalez, A. Maritan, and A.-L. Barabasi, "A universal model for mobility and migration patterns," *Nature*, vol. 484, no. 7392, pp. 96–100, 04 2012.

[15] J.-P. Onnela, S. Arbesman, M. C. González, A.-L. Barabási, and N. A. Christakis, "Geographic constraints on social network groups," *PLoS ONE*, vol. 6, no. 4, p. e16939, 04 2011. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pone.0016939

[16] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi, "Human mobility, social ties, and link prediction," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '11. New York, NY, USA: ACM, 2011, pp. 1100–1108.

[17] R. Becker, R. Cáceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky, "Human mobility characterization from cellular network data," *Commun. ACM*, vol. 56, no. 1, pp. 74–82, Jan. 2013.

[18] J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási, "Uncovering individual and collective human dynamics from mobile phone records," *Journal of Physics A: Mathematical and Theoretical*, vol. 41, no. 22, p. 224015, 2008.

[19] J. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, and A. L. Barabasi, "Structure and tie strengths in mobile communication networks," *Proc. Natl. Acad. Sci. USA*, vol. 104, no. 18, pp. 7332–7336, 2007.

[20] M. González, C. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, June 2008.

[21] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil, "An empirical study of geographic user activity patterns in foursquare," in *International AAAI Conference on Weblogs and Social Media*, 2011.

[22] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo, "A tale of many cities: universal patterns in human urban mobility," *PloS one*, vol. 7, no. 5, p. e37027, 2012.

[23] N. Li and G. Chen, "Analysis of a location-based social network," in *Computational Science and Engineering, 2009. CSE '09. International Conference on*, vol. 4, Aug 2009, pp. 263–270.

[24] M. Allamanis, S. Scellato, and C. Mascolo, "Evolution of a location-based online social network: Analysis and models," in *Proceedings of the 2012 ACM Conference on Internet Measurement Conference*, ser. IMC '12. New York, NY, USA: ACM, 2012, pp. 145–158.

[25] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil, "Exploiting semantic annotations for clustering geographic areas and users in location-based social networks," in *Proceedings of 3rd Workshop Social Mobile Web (SMW'11)*, Barcelona, Spain, Jul. 2011.

[26] S. Scellato and C. Mascolo, "Measuring user activity on an online location-based social network," in *Proceedings of Third International Workshop on Network Science for Communication Networks (NetSciCom)'11*, Shanghai, China, Apr. 2011.

[27] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '11. New York, NY, USA: ACM, 2011, pp. 1082–1090.

[28] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui, "Exploring millions of footprints in location sharing services." in *ICWSM*. The AAAI Press, 2011.

[29] H. Gao, J. Tang, and H. Liu, "Exploring social-historical ties on location-based social networks," in *ICWSM*, 2012.

[30] M. A. Vasconcelos, S. Ricci, J. Almeida, F. Benevenuto, and V. Almeida, "Tips, dones and todos: Uncovering user profiles in foursquare," in *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, ser. WSDM '12. New York, NY, USA: ACM, 2012, pp. 653–662.

[31] L. Ferrari, A. Rosi, M. Mamei, and F. Zambonelli, "Extracting urban patterns from location-based social networks," in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, ser. LBSN '11. New York, NY, USA: ACM, 2011, pp. 9–16.

[32] L. Ferrari and M. Mamei, "Discovering daily routines from google latitude with topic models," in *PerCom Workshops*. IEEE, 2011, pp. 432–437.

[33] K. Farrahi and D. Gatica-Perez, "Discovering routines from large-scale human locations using probabilistic topic models," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 1, pp. 3:1–3:27, Jan. 2011.

[34] N. Eagle, A. Pentland, and D. Lazer, "Inferring social network structure using mobile phone data," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 106, p. 1527415278, 2009.

[35] (2008) 10 emerging technologies 2008. [Online]. Available: http://www.technologyreview.com/specialreports/specialreport.aspx?id=25.

[36] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and pois," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '12. New York, NY, USA: ACM, 2012, pp. 186–194.

[37] J. Cranshaw and T. Yano, "Seeing a home away from the home: Distilling proto-neighborhoods from incidental data with latent topic modeling," in *NIPS'10 Workshop of Computational Social Science and the Wisdom of the Crowds*, 2010.

[38] J. Chang and E. Sun, "Location3: How Users Share and Respond to Location-Based Data on Social Networking Sites," in *Proceedings of the Fifth International Conference on Weblogs and Social Media*. AAAI, Jul. 2011.

[39] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from gps trajectories," in *Proceedings of the 18th International Conference on World Wide Web*, ser. WWW '09. New York, NY, USA: ACM, 2009, pp. 791–800.

[40] Y. Zheng and X. Xie, "Learning travel recommendations from user-generated gps traces," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 1, pp. 2:1–2:29, Jan. 2011.

[41] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang, "Collaborative location and activity recommendations with gps history data," in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 1029–1038.

[42] K. Joseph, C. H. Tan, and K. M. Carley, "Beyond "local", "categories" and "friends": Clustering foursquare users with latent "topics"," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, ser. UbiComp '12. New York, NY, USA: ACM, 2012, pp. 919–926.

[43] X. Long, L. Jin, and J. Joshi, "Exploring trajectory-driven local geographic topics in foursquare," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, ser. UbiComp '12. New York, NY, USA: ACM, 2012, pp. 927–934.

[44] W. Sherchan, S. Nepal, and C. Paris, "A survey of trust in social networks," *ACM Comput. Surv.*, vol. 45, no. 4, pp. 47:1–47:33, Aug. 2013.

[45] J. Caverlee, L. Liu, and S. Webb, "Socialtrust: Tamper-resilient trust establishment in online communities," in *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*, ser. JCDL '08. New York, NY, USA: ACM, 2008, pp. 104–114.

[46] J. Golbeck and J. Hendler, "Inferring binary trust relationships in web-based social networks," *ACM Trans. Internet Technol.*, vol. 6, no. 4, pp. 497–529, Nov. 2006.

[47] J. A. Golbeck, "Computing and applying trust in web-based social networks," Ph.D. dissertation, University of Maryland at College Park, College Park, MD, USA, 2005.

[48] C.-N. Ziegler and G. Lausen, "Spreading activation models for trust propagation," in *Proceedings of the 2004 IEEE International Conference on e-Technology, e-Commerce and e-Service (EEE'04)*, ser. EEE '04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 83–97.

[49] M. Jamali and M. Ester, "Trustwalker: A random walk model for combining trust-based and item-based recommendation," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '09. New York, NY, USA: ACM, 2009, pp. 397–406.

[50] C. Wilson, B. Boe, A. Sala, K. P. Puttaswamy, and B. Y. Zhao, "User interactions in social networks and their implications," in *Proceedings of the 4th ACM European Conference on Computer Systems*, ser. EuroSys '09. New York, NY, USA: ACM, 2009, pp. 205–218.

[51] S. Adali, R. Escriva, M. K. Goldberg, M. Hayvanovych, M. Magdon-Ismail, B. K. Szymanski, W. A. Wallace, and G. T. Williams, "Measuring behavioral trust in social networks," in *IEEE International Conference on Intelligence and Security Informatics(ISI)*, 2010, pp. 150–152.

[52] H. Liu, E.-P. Lim, H. W. Lauw, M.-T. Le, A. Sun, J. Srivastava, and Y. A. Kim, "Predicting trusts among users of online communities: An epinions case study," in *Proceedings of the 9th ACM Conference on Electronic Commerce*, ser. EC '08. New York, NY, USA: ACM, 2008, pp. 310–319.

[53] S. Nepal, W. Sherchan, and C. Paris, "Strust: A trust model for social networks," in *Proceedings of the 2011IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications*, ser. TRUSTCOM '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 841–846.

[54] S. Trifunovic, F. Legendre, and C. Anastasiades, "Social trust in opportunistic networks," in *INFOCOM IEEE Conference on Computer Communications Workshops , 2010*, March 2010, pp. 1–6.

[55] M. Ye, P. Yin, and W.-C. Lee, "Location recommendation for location-based social networks," in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ser. GIS '10. New York, NY, USA: ACM, 2010, pp. 458–461.

[56] M. Ye, P. Yin, W.-C. Lee, and D.-L. Lee, "Exploiting geographical influence for collaborative point-of-interest recommendation," in *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '11. New York, NY, USA: ACM, 2011, pp. 325–334.

[57] K. W.-T. Leung, D. L. Lee, and W.-C. Lee, "Clr: A collaborative location recommendation framework based on co-clustering," in *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '11. New York, NY, USA: ACM, 2011, pp. 305–314.

[58] Y. Zheng, L. Zhang, Z. Ma, X. Xie, and W.-Y. Ma, "Recommending friends and locations based on individual location history," *ACM Trans. Web*, vol. 5, no. 1, pp. 5:1–5:44, Feb. 2011.

[59] J. Bao, Y. Zheng, and M. F. Mokbel, "Location-based and preference-aware recommendation using sparse geo-social networking data," in *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, ser. SIGSPATIAL '12. New York, NY, USA: ACM, 2012, pp. 199–208.

[60] J. J.-C. Ying, E. H.-C. Lu, W.-N. Kuo, and V. S. Tseng, "Urban point-of-interest recommendation by mining user check-in behaviors," in *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, ser. UrbComp '12. New York, NY, USA: ACM, 2012, pp. 63–70.

[61] S. Scellato, A. Noulas, and C. Mascolo, "Exploiting place features in link prediction on location-based social networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '11. New York, NY, USA: ACM, 2011, pp. 1046–1054.

[62] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo, "Socio-spatial properties of online location-based social networks," in *International AAAI Conference on Weblogs and Social Media*, 2011.

[63] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo, "Mining user mobility features for next place prediction in location-based services," in *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*, ser. ICDM '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 1038–1043.

[64] Foursquare api. [Online]. Available: https://developer.foursquare.com/index

[65] Long tail. [Online]. Available: http://en.wikipedia.org/wiki/Long_Tail

[66] H. Zang and J. Bolot, "Anonymization of location data does not work: A large-scale measurement study," in *Proceedings of the 17th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '11. New York, NY, USA: ACM, 2011, pp. 145–156.

[67] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by latent semantic analysis," *JASIS*, vol. 41, no. 6, pp. 391–407, 1990.

[68] J. Cranshaw, R. Schwartz, J. Hong, and N. Sadeh, "The livehoods project: Utilizing social media to understand the dynamics of a city," in *ICWSM'12*, 2012.

[69] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. Suppl. 1, pp. 5228–5235, April 2004.

[70] http://www.ticketmaster.com/promo/u3c7b3?brand=\penguins.

[71] F. Kling and A. Pozdnoukhov, "When a city tells a story: Urban topic analysis," in *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, ser. SIGSPATIAL '12. New York, NY, USA: ACM, 2012, pp. 482–485.

[72] C. Brown, V. Nicosia, S. Scellato, A. Noulas, and C. Mascolo, "The importance of being placefriends: discovering location-focused online communities," in *WOSN*, 2012, pp. 31–36.

[73] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg, "Inferring social ties from geographic coincidences," *Proceedings of the National Academy of Sciences*, vol. 107, no. 52, pp. 22 436–22 441, Dec. 2010.

[74] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, pp. 38:1–38:55, Sep. 2014.

[75] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proceedings of the 10th International Conference on World Wide Web*, ser. WWW '01. New York, NY, USA: ACM, 2001, pp. 285–295.

[76] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: Item-to-item collaborative filtering," *IEEE Internet Computing*, vol. 7, no. 1, pp. 76–80, Jan. 2003.

[77] M. Deshpande and G. Karypis, "Item-based top-n recommendation algorithms," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 143–177, Jan. 2004.

[78] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, Aug. 2009.

[79] M. Jamali and M. Ester, "A matrix factorization technique with trust propagation for recommendation in social networks," in *Proceedings of the Fourth ACM Conference on Recommender Systems*, ser. RecSys '10. New York, NY, USA: ACM, 2010, pp. 135–142.

[80] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Trans. on Knowl. and Data Eng.*, vol. 17, no. 6, pp. 734–749, Jun. 2005.

[81] D. Almazro, G. Shahatah, L. Albdulkarim, M. Kherees, R. Martinez, and W. Nzoukou, "A survey paper on recommender systems," Arxiv preprint, 2010. [Online]. Available: http://arxiv.org/abs/1006.5278v4

[82] H. Kautz, B. Selman, and M. Shah, "Referral web: Combining social networks and collaborative filtering," *Commun. ACM*, vol. 40, no. 3, pp. 63–65, Mar. 1997.

[83] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri, "Feedback effects between similarity and social influence in online communities," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '08. New York, NY, USA: ACM, 2008, pp. 160–168.

[84] J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy, "Make new friends, but keep the old: Recommending people on social networking sites," in *Proceedings of the SIGCHI*

Conference on Human Factors in Computing Systems, ser. CHI '09. New York, NY, USA: ACM, 2009, pp. 201–210.

[85] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, no. 5, pp. 604–632, Sep. 1999.

[86] S. Ihara, *Information theory for continuous systems*. World Scientific, 1993.

[87] K. Pelechrinis and P. Krishnamurthy, "Location-based social network users through a lense: Examining temporal user patterns," in *AAAI Fall Symposium Series*, 2012.

[88] H. Tong, C. Faloutsos, and J.-Y. Pan, "Fast random walk with restart and its applications," in *Proceedings of the Sixth International Conference on Data Mining*, ser. ICDM '06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 613–622.