Chapter 8
Data Analysis on Location-Based Social Networks

Huiji Gao and Huan Liu

Abstract The rapid growth of location-based social networks (LBSNs) has greatly enriched people's urban experience through social media, and attracted increasing number of users in recent years. Typical location-based social networking sites allow users to "check in" at a physical place and share the location with their online friends, and therefore bridge the gap between the real world and online social networks. The availability of large amounts of geographical and social data on LBSNs provides an unprecedented opportunity to study human mobile behavior through data analysis in a spatial–temporal–social context, enabling a variety of location-based services, from mobile marketing to disaster relief. In this chapter, we first introduce the background and framework of location-based mobile social networking. We next discuss the distinct properties, data analysis and research issues of location-based social networks, and present two illustrative examples to show the application of data mining to real-world location-based social networks.

8.1 Introduction

The wide use of mobile devices and location-based services in the world has generated a new concept of online social media, namely location-based social networks (LBSNs). Location-based social networking sites use GPS, Web 2.0 technology and mobile devices to allow people to share their locations (usually referred to as "check-in"), find out local points of interest and discounts, leave comments on specific places, connect with their friends, and find other friends who are nearby. A recent survey from the Pew Internet and American Life Project reports that over the past year, smartphone ownership among American adults has risen from 35 %

H. Gao (✉) • H. Liu
Computer Science and Engineering, Arizona State University, Phoenix, USA
e-mail: Huiji.Gao@asu.edu; Huan.Liu@asu.edu

A. Chin and D. Zhang (eds.), *Mobile Social Networking: An Innovative Approach*, Computational Social Sciences, DOI 10.1007/978-1-4614-8579-7_8,

in 2011 to 46 % in 2012. Almost three-quarters (74 %) of smartphone owners use their phone to get real-time location-based information such as getting directions or recommendations. Meanwhile, 18 % of smartphone owners use geo-social services, such as Foursquare,[1] Gowalla,[2] and Facebook Places,[3] to "check in" to certain locations and share them with their friends, this percentage having risen from 12 % in 2011 (Zickuhr 2012). It is anticipated that more than 82 million users will subscribe to location-based social networking services by 2013 (ABI Research 2008), and location-based marketing will be a $1.8 billion business worldwide by 2015 (ABI Research 2010). Such rapid growth of location-based social networks has led to the availability of a large amount of user data, which consists of both the geographical trajectories and the social friendships of users, providing both opportunities and challenges for researchers to investigate users' mobile behavior in spatial, temporal, and social aspects.

Typical online location-based social networking sites provide location-based services that allow users to "check in" at physical places, and automatically include the location into their posts. "Check-in" is an online activity that posts a user's current geographical location to tell his friends when and where he is through social media. Compared with many other online activities (following, grouping, voting, tagging, etc.) that interact with the virtual world, "check-in" reflects a user's geographical action in the real world, residing where the online world and real world intersect. In this scenario, "check-in" not only adds a spatial dimension to the online social networks, but also plays an important role in bridging the gap between the real world and the virtual world. Thus, the study of check-ins on location-based social networks provides an ideal environment to analyze users' real world behavior through virtual media, and could potentially improve a variety of location-based services such as mobile marketing (Barnes and Scornavacca 2004; Bauer et al. 2005; Scharl et al. 2005), disaster relief (Goodchild and Glennon 2010; Gao et al. 2011a, b), and traffic forecasting (Ben-Akiva et al. 1998; Dia 2001).

The first commercial location-based social networking service available in the United States is Dodgeball,[4] launched in 2000. It allows users to "check in" by broadcasting their current locations through short messages to their friends who are within a ten-block radius; users can also send "shouts" to organize a meeting among friends at a specific place. After being acquired by Google in 2005, the original Dodgeball was replaced with Google Latitude in 2009, while the founder of Dodgeball launched a new location-based social networking service, "Foursquare", in the same year. Foursquare utilizes a game mechanism in which users can compete for virtual positions, such as mayor of a city, based on their check-in activities. It reached 20 million users by April 2012 (Kessler 2012), becoming one of the most successful location-based social networking sites in the United States. Facebook

[1] http://foursquare.com

[2] http://en.wikipedia.org/wiki/Gowalla

[3] http://www.facebook.com/about/location

[4] http://en.wikipedia.org/wiki/Dodgeball

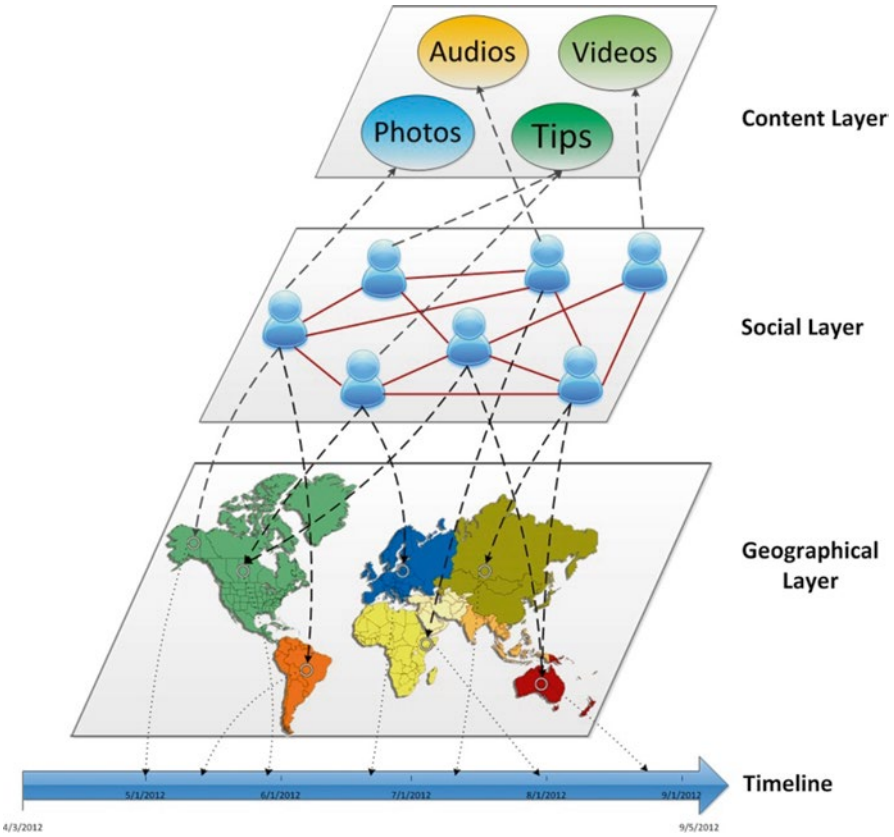

Fig. 8.1 The information layout of location-based social networks

also launched its location-based service, namely Facebook Places, in 2010 with its check-in function, and acquired another popular LBSN, Gowalla,[5] at the end of 2011. All these location-based social networking sites share a "3 + 1" framework, i.e., three layers and one timeline, as shown in Fig. 8.1.

The geographical layer contains the historical check-ins of users, while the social layer contains social friendship information, and the content layer consists of user feedbacks or tips about different places. All these three layers share one timeline, indicating the temporal information of the user "check-in" behavior. Previous research has investigated the social and content layers with traditional online social network data (Hu and Liu 2012), and analyzed the geographical and content layers with mobile phone data (Chen and Kotz 2000). Compared to them, location-based social network data has an additional geographical layer which is not available in traditional online social networks, and an explicit social layer which is not available

[5] http://www.pcmag.com/article2/0,2817,2401433,00.asp

from mobile phone data (usually social friendship information from mobile phone data is derived through smartphone proximity network). The unique geographical property and the social network information presents new challenges for data analysis on location-based social network data, since traditional approaches on social network or mobile phone data may fail due to the lack of pertinence. Furthermore, the "3+1" data structure defines six different types of networks, i.e., location–location network, user–user network, content–content network (e.g., word–word network), user-location network, user-content network, and location-content network. Each one can be mined together with the temporal information provided by the timeline, indicating more opportunities for data analysis on LBSNs. Therefore, data analysis techniques specifically designed for LBSNs can efficiently deal with these distinct properties, and help understand user behavior for research and business purposes.

The rest of this chapter is organized as follows. We first introduce the distinct properties of location-based social network data in Sect. 8.2, then discuss the data analysis and research issues in Sect. 8.3, followed by two real-world examples of applying data mining to location-based social networks in Sect. 8.4, and finally provide some conclusions with suggestions for future work in Sect. 8.5.

8.2 Distinct Properties of Location-Based Social Network Data

Location-based social networks provide data consisting of both geographical information and social networks. Compared to traditional online social network data and mobile phone data, location-based social network data have distinct properties in several aspects.

8.2.1 Geographical Property

One of the most significant differences between LBSNs and traditional online social networks is the geographical property, which is considered as the unique facet of location-based social networks. Users on LBSNs are able to check in at a physical place, and let their friends be aware of this check-in. The check-in location indicates the current geographical status of a user in the real world, and generates the local social networks of the user based on this location. In this scenario, the geographical check-in locations bridge the gap between the real world and online social networks (Cranshaw et al. 2010; Gao et al. 2012a), which in turn reflect the user's behavior more closely to the real world compared with other online social networks, and provide an unprecedented opportunity to study a user's real-world behavior through social media. Researchers have studied the distinctions between online and offline social networks (Cranshaw et al. 2010), differences between location-based social networks and content-based social networks (Scellato et al. 2010), and relationship between geographical

distance and friendship (Scellato et al. 2011b; Cho et al. 2011), etc. These analyses exploit many fundamental user mobile patterns, and motivate us to make use of geographical properties for the development of better location-based services.

1. *Large-Scale Mobile Data*
   The increasing use of mobile devices and popular location-based mobile social networking sites has led to the massive availability of mobile data. Compared with the traditional cell phone data, which is usually collected through telecommunication carriers with limited number of users (Zheng et al. 2009), location-based social networking services utilize Web 2.0 technology combined with GPS on mobile devices, generating a large amount of geographical and social information from millions of users (Chang and Sun 2011; Scellato et al. 2011b). For example, Google Latitude reported ten million active users in 2011,[6] Yelp had approximately 71 million unique visitors monthly on average in the first quarter of 2012,[7] and Foursquare reached 20 million users and two billion check-ins by April 2012 (Kessler 2012). Researchers can easily obtain these data through public APIs provided by location-based social networking sites, enabling the large-scale data analysis of user behavior in a spatial, temporal, and social context (Cheng et al. 2011; Scellato et al. 2011b; Gao et al. 2012a).
2. *Accurate Description of Geolocations*
   Location-based mobile social networking sites provide more accurate location descriptions than traditional geo-tagged data. For example, in location-based social networks, it is easy to distinguish two adjacent restaurants on a street, two nearby stores in a fashion square, or a pharmacy located upstairs of a bar. This is because the traditional geo-tagged data only provide the longitude and latitude of a location, while location-based social networking sites such as Foursquare and Facebook places could provide additional textual descriptions for popular venues, e.g., categories, comments, and tips, therefore promoting a variety of location-based applications from location recommendation (Ye et al. 2011a) to urban computing (Cranshaw et al. 2012) by endowing the physical places with semantic meaning.
3. *Data Sparseness*
   In traditional cell phone data, a user's geographical location is automatically recorded by the telecommunication tower, while on location-based social networks, the check-in process is user-driven (Noulas et al. 2011a), i.e., the user decides whether to check in at a specific place or not due to certain privacy concerns. For example, a user may usually check in at Starbucks in New York, but with the latest check-in at SeaWorld in San Diego, or check in continuously at the same restaurant many times. Some users even have more than 1-year gaps between consecutive check-ins. Such check-in behavior leads to the significant sparseness of geographical data in location-based social networks, which greatly increases the difficulty of data analysis, especially in investigating human mobility patterns.

[6] http://techcrunch.com/2011/02/01/google-latitude-check-in

[7] http://www.yelp-press.com/phoenix.zhtml?c=250809&p=irol-press

4. *Explicit Social Friendship*
   The social networks on location-based social networking sites consist of social friendship information explicitly defined by users (a user can explicitly add another user as a friend), while in traditional cell phone data, the social network is usually collected through user study (Li et al. 2008; Eagle et al. 2009), or derived from communication network or Bluetooth network (Wang et al. 2011). This property enables more accurate and efficient data analysis and evaluation on location-based social networks, especially for applications such as friend recommendation and location privacy control (Kelley et al. 2008).

8.3 Data Analysis and Research Issues of Location-Based Mobile Social Networks

The heterogeneous data in location-based social networks contain spatial–temporal–social context and present new challenges and opportunities for data analysis. One can ask many interesting questions that can potentially be answered by analyzing LBSN data. For example, are there any relationships between user attitudes and mobile patterns on LBSNs? How does geographical distance affect online social friendship, and vice versa? Why do people use location-based social networking services? Under what circumstances would users not like to share their locations due to privacy issues? Can location prediction help mobile marketing? Can location-recommender systems improve urban experience? How can one best control location privacy to maximize her social networking experience? In this section, we introduce a variety of data analysis techniques and current research on location-based social networks, and show how answers to these challenging questions can be obtained via novel data analysis to improve location-based services.

8.3.1 Social Friendship and Geographical Distance

Traditional social networking analysis mainly studies network structure and properties, which does not consider the geographical distance between nodes. In 2001, Cairncross (2001) proposed the term "the death of distance", claiming that geographical distance begins to play a less important role due to the communication revolution and the rapid development of the Internet, which therefore could lead our world to a "global village". Later, Gastner and Newman (2006) studied the spatial structure networks. They demonstrated that there is a strong correlation between geographical attributes and network properties, indicating the significance of considering the spatial properties of networks for future applications. Other researchers studied geographical distance in the Internet, and argued that the IT revolution does not transfer us into a borderless society, as physical proximity still plays an

important role in the Internet era (Goldenberg and Levy 2009; Mok et al. 2010). All these studies are based on traditional networks such as e-mail networks, cell phone contact networks, road networks, and the Internet.

One of the first attempts to investigate how social connection is affected by geographical distance in online social networks was proposed by Liben-Nowell et al. (2005). The authors studied users' social networks and their hometown information obtained from LiveJournal. Their simulation model shows that one-third of friendships are independent of geography. With the wide use of mobile devices, such as Apple iPhones and Google Android phones, and the increasing attention on mobile social networking, location-based social networks focused on the small local social network derived from a user's geographical location become more and more popular. Dodgeball was the first commercial location-based social network service available in the United States, launched in 2000. Humphreys (2007) studied user behavior on Dodgeball, and found that LBSNs do change people's attitude toward locations and their experience of urban life.

The increasing popularity of location-based social networking sites makes it possible to obtain data consisting of the geographical distance between users and their social networks in large-scale, which in turn enables a vast research opportunity for large-scale data analysis on geo-social properties in LBSNs. Scellato et al. (2010) proposed two geo-social metrics, embedding the geographical distance into social structure, to measure the node locality and geographical clustering coefficient. Two findings are presented in this work: (1) users who live close have a higher probability to create friendship links than those who live at a distance, and (2) users in the same social cluster show short geographical distances. Furthermore, the authors compared location-based social networks (Brightkite and Foursquare) with content-sharing-based social networks (LiveJournal and Twitter), discovering the difference of network properties between these two kinds of social networks. They found that people within a social cluster on the LBSNs tend to have smaller geographical distance than those online social networks focusing on content producing and sharing.

Researchers have also investigated how geographical distance influences social networks, and how social networks influence human movement on LBSNs. Scellato et al. (2011b) presented a comprehensive study on three location-based social networking sites, i.e., Brightkite, Foursquare, and Gowalla. They observed strong heterogeneity across users with different geographic scales of interaction across social ties, with the probability of a social tie between two users as a function of the geographical distance between them. Cho et al. (2011) studied Gowalla, Brightkite, and cell phone data, reporting that long-distance travel is more influenced by social friendship, while short-range human movement is not influenced by social networks. More recently, Kulshrestha et al. (2012) investigated the Twitter social network, and concluded that offline geography still matters in online social networks, while one-third of the users would like to have their social links in other countries, which is consistent with the previous findings presented in Liben-Nowell et al. (2005) and Scellato et al. (2010). Brown et al. (2012) extended the research on LBSNs to social community, and discovered that the rise of social groups is affected

by both social and spatial factors. They reported that social communities on location-based social networks seem to be more relevant to the spatial factor. This is also consistent with previous findings (Scellato et al. 2010) about the differences between location-based social networks and content-sharing-based social networks.

8.3.2 User Activity and Mobile Pattern Analysis

Sociologists have studied the characteristics of user behavior on location-based social networks, motivated by the potential power of these characteristics for future research and applications. Among the current research, there are two major characteristics that sociologists mostly discussed, i.e., user activity and mobile patterns.

1. *User Activity*
   User activity indicates how frequently a user creates and consumes online content in LBSNs. Researchers attempt to classify users into various groups, representing different levels of user activity. This is motivated by tailoring location-based services to different user types to benefit the majority of users. One of the first large-scale analyses of user activity on a real-world commercial location-based social network was presented in Li and Chen (2009). The authors analyzed user profiles on Brightkite, and observed that the majority of users are male users who are professionals and willing to participate in social media. They also found that users with higher network degree tend to be more mobile and active. The authors further clustered users based on their attributes such as total number of updates, uniquely visited places, etc., and obtained five user groups according to user activity, named as inactive, normal, active, mobile, and trial users. They reported that the majority of users on Brightkite are trial users, while only 6 % of users are clustered as active users. Noulas et al. (2011b) used a spectral clustering algorithm to group users based on their check-in category distribution on Foursquare, aiming at identifying user communities to help develop new applications such as recommender systems.

   Vasconcelos et al. (2012) considered different type of features for user clustering on Foursquare. They focused on the tips, dones, and to-dos of venues, and utilized three related attributes to cluster users, i.e., the number of tipped venues, the total number of dones and to-dos, and the percentage of tips with links. They obtained four groups, with three groups based on user activity level, and one group representing spam users. It is reported that around 86 % of users tend to tip a larger number of venues and get more dones and to-dos in return, forming the largest group on Foursquare. Furthermore, the authors showed that observing a large number of links pointed to unrelated content in tips can be a good predictor for detecting spam users.
2. *Mobile Patterns*
   Cheng et al. (2011) explored millions of check-ins on Facebook, and observed various spatial, temporal, and social patterns. For example, human movement follows a "Lévy Flight" (Rhee et al. 2011), in which people tend to move to nearby places and occasionally to distant places. The authors observed that user

mobility is influenced by social status, geographical, and economic factors. Furthermore, the user check-in behavior presents strong daily/weekly patterns and periodic property, indicating the potential to improve location-based applications. In Noulas et al. (2011a), the authors observed similar geo-temporal patterns of check-ins on weekdays and weekends. They reported that around 20 % of consecutive check-ins in Foursquare happen within 1 km of one another, 60 % between 1 and 10 km, and 20 % over 10 km. Li and Chen (2009) studied users' mobility characteristics on Brightkite. They clustered users based on their mobility patterns derived from user updates and movement paths, and obtained four user groups, namely home users, home–vacation users, home–work users, and other users which present different mobility patterns from previous groups.

8.3.3 Location Prediction

Location prediction is a traditional task in mobile computing. It has been studied over a long period. Researchers analyze human mobility patterns to improve location prediction services, and therefore exploit their potential power on various applications such as mobile marketing (Barnes and Scornavacca 2004; Barwise and Strong 2002), traffic planning (Ben-Akiva et al. 1998; Dia 2001), and even disaster relief (Gao et al. 2011a; Goodchild and Glennon 2010; Gao et al. 2012a; Wang and Huang 2010). Current research on location prediction in LBSNs mainly focuses on two tasks: (1) predicting a user's home location, and (2) predicting a user's location at any time. The former task considers the static home location of a user, while the latter considers more about a user's moving trajectories, with his location in movement.

Before we delve into different location prediction methods, we first discuss two commonly used evaluation metrics in the location prediction task. The first metric is *prediction accuracy*, i.e., the fraction of correctly predicted locations over the total number of predicted locations in the testing set, which has been widely used in current work (Gao et al. 2012a; Cho et al. 2011; Backstrom et al. 2010). Sometimes its variants have also been used for additional evaluation. For example, the top-k accuracy is utilized in Cheng et al. (2010). It returns the top k candidates as the predictions for a location, and treats a prediction as correct as long as the ground truth location is among the top k returned locations. Here, k is usually selected as 2, 3, 5, and 10. The second metric is *expected distance error* (Cho et al. 2011), as shown below, which computes the average geographical distance between the real location and the estimated location, over all predicted locations.

$$ErrD = \frac{1}{|L|}\sum_{lPL} d\left(l_{act}, l_{est}\right) \tag{8.1}$$

where L is the unknown locations in the testing set, l_{act} is the actual location, and l_{est} is the estimated location. d(x,y) is a function that computes the geographical distance between two locations x and y.

The motivation of home location prediction arises from the sparseness of available user home locations on popular social networks such as Twitter and Facebook. Based on the statistics from Cheng et al. (2010), only 26 % of Twitter users list their locations as granularly as a city name, and less than 0.42 % of all tweets use the geo-tagging function to indicate their locations. On the other hand, the availability of user home location leads to a user-centric social network. It provides an opportunity to study social networks from a user's ego view, and in turn benefits applications such as targeting advertisement regions, and summarizing the local news for nearby users. Therefore, obtaining the user home location is critical to studying human mobility on location-based social networks.

Current work in home location prediction on LBSNs uses two kinds of resources, i.e., content information and social network information. The content-based approaches (Cheng et al. 2010; Hecht et al. 2011) studied the location information implicated in a user's tweet content, and proposed a location prediction framework based on the correlation between specific terms in tweets and their corresponding locations.

Backstrom et al. (2010) utilized social network information on Facebook to predict the user's home location. They predicted a Facebook user's home address based on the provided home addresses of his friends. One observation was leveraged so that the probability of a link being present between two nodes is a function of their geographical distance. By maximizing the likelihood of observations on friendship and non-friendship of a user, the unknown home location could be computed according to friends' addresses. All these methods predict the location at country, state, or city level, while the spatial resolution is low.

To predict a user's location at any time, usually referred to as *next location prediction*, various approaches have been proposed in the last decade. Without the social network information being available, these methods mainly consider the spatial trajectories (Monreale et al. 2009; Spaccapietra et al. 2008), temporal patterns (Thanh and Phuong 2007), or spatial–temporal patterns (Scellato et al. 2011a; Gao et al. 2012c) for location prediction. With the availability of social information on LBSNs, Gao et al. (2012a) proposed the first work of modeling social information for next location prediction on LBSNs with a social-historical model. Later, Noulas et al. (2012) further investigated the next location prediction problem and proposed a set of features regarding various facets of user behavior for prediction. Researchers have made a great effort to investigate the role of social friendship in explaining a user's mobile patterns. On the other hand, leveraging social networking information for location prediction becomes a new challenge, since how to embed the social property into geographical patterns is still an open issue on location-based social networks (Gao et al. 2012b).

Current work on LBSNs has proposed various approaches to combing social network information with traditional spatial–temporal patterns. Chang and Sun (2010) utilized logistic regression model to combine a set of features extracted from Facebook data. The features include a user's previous check-ins, user's friends' check-ins, demographic data, distance of place to user's usual location, etc. Their results demonstrated that the number of previous check-ins by the user is a strong

predictor, while previous check-ins made by friends and the age of the user are also good features for prediction.

Linear combination has been mostly used for integrating social friendship with spatial–temporal patterns (Cho et al. 2011; Gao et al. 2012a). Cho et al. (2011) considered the user check-in probability as a linear combination of social effect and non-social effect. The social effect assumes the check-in of a user to be close to the check-ins of his friends, both in space and in time; while the non-social effect captures the periodical patterns, which considers the user's personal movement following a 2-D Gaussian distribution, with the two Gaussian centers focusing on home and work. Gao et al. (2012a) proposed a social–historical model integrating the social ties and historical ties of a user for location prediction. Both ties generate the probability of next location based on the observation of previous check-in sequence. The historical ties consider the user's own check-in sequence, and the social ties consider the check-in sequences of the user's friends. Based on the observation that word sequence and location trajectory share a set of common properties, a language model is then introduced for generating the next location probability.

All of the current work reports very limited improvement by utilizing social network information in LBSNs. The model that considers social networks slightly improves those that do not consider social networks. However, this does not lead to the conclusion that social network has no contributions to a user's mobility. The best way to integrate the social network and leverage it for location prediction is still under study.

8.3.4 Recommender Systems

Recommender systems are designed to recommend items to users in various situations such as online shopping, dating, and social events. Since the exploration of city and neighborhood provides us with more choices of life experience than before, recommendation is indispensable to help users filter uninteresting items, and therefore reduce their time in decision-making. Furthermore, recommender systems could also benefit virtual marketing, since the appropriate recommendations could attract users with specific interests. Recommender systems on location-based social networks only started just a few years ago, and three items are mainly recommended in current work, which are locations, tags, and friends.

1. *Location Recommendation*
   Location recommendation aims to recommend a set of locations to a user based on the user's interests. The major difference between location prediction and location recommendation is that location prediction usually predicts the next location as an existing location that the user has been before, while location recommendation would recommend a new location that the user has never been before. From a research standpoint, location prediction on LBSNs considers more how to utilize the social information, while current research in location

recommendation on LBSNs mainly focuses on the geo-spatial and temporal influence, and the social network information is usually utilized through traditional collaborative filtering (Berjani and Strufe 2011; Zhou et al. 2012), which considers the location as an item such as that on Epinions (Tang et al. 2012a, b). For evaluation, performance@N (Ye et al. 2011c) is usually adopted to evaluate the location recommendation performance. The performance@N metric consists of precision@N and recall@N. It consider all the locations that should be recommended as uncovered locations, and the set of correctly recommended locations as recovered locations. The precision@N evaluates the ratio of recovered locations to the N recommended locations, and the recall@N calculates the ratio of recovered locations to uncovered locations.

Ye et al. (2010) first introduced location recommendation on location-based social networks. In this paper, the major focus is location recommendation efficiency. The essential content contains: (1) only friendship information was used for collaborative filtering, and (2) instead of calculating the user similarity based on historical behavior (e.g., check-in history), the authors captured the correlations between geographical distance and user similarity, and leveraged them for user similarity calculation. This work is later extended in Ye et al. (2011c), which considers both spatial influence and social friendships for location recommendation. Three factors are investigated and combined to recommend locations. The first factor represents influence from similar users, the second factor indicates influence from friends, and the third factor captures geographical influence, under the hypothesis that people tend to visit close places more often than distant places. A spatial constraint is generated to capture the geographical influence by exploiting the relationship between a user visiting two places and the geographical distance between these two places. These three factors are then represented by three probabilities, and linearly combined together with corresponding weights. The results demonstrated that the most influential factor actually comes from the similar users, while friendship and geographical distance together have around 30 % influences.

2. *Tag Recommendation*
   Tag recommendation is motivated to enrich the semantic meaning of places and to facilitate the development of recommender systems such as "Point of Interest" retrieval services. Temporal patterns have been usually considered for tag recommendation on location-based social networks. In Ye et al. (2011a), the authors proposed "temporal bands" to capture the temporal patterns of each place, and suggested their potential ability for tag recommendation. For example, a bar may be visited frequently at 11:00 p.m. to 1:00 a.m., while a restaurant may have more visits around 12:00 p.m. and 6:00 p.m. Therefore, tags associated with the bar or restaurant present different visiting distributions over time, i.e., temporal bands. By considering the visiting probability at different hours of a day and different days of a week, one can compare such visiting distributions between candidate tags and target places; the recommender system could then recommend a set of tags that mostly fit the temporal band of that place. In this work, the authors only proposed the idea of temporal bands, but did not apply it to real-world datasets for tag recommendation.

In Ye et al. (2011b), the temporal information has been formally utilized for tag recommendation and place annotation. In this work, the authors considered tag recommendation as a classification problem. Two sets of features, named explicit patterns and implicit patterns, are firstly defined to generate the feature space for each place, then a SVM classifier is learned for each tag, based on the observed feature vectors that are associated and not associated with the tag. The explicit patterns include features that can be explicitly observed in the data, e.g., total number of check-ins, total number of unique visitors, etc. The implicit patterns generate the relatedness between two places based on their common visiting users and common temporal patterns, while the latter factor is similar to Ye et al. (2011a). These two factors are linearly combined together, which generates a ranking list of places based on their relatedness to the target place. A place with high relatedness is referred to as a semantic neighbor, and the corresponding relatedness indicates the probability of the target place to be labeled with a given semantic tag from this neighbor. The final implicit patterns are the probabilities for each possible tag on the target place. The hypothesis of this method is that two places checked in by the same user around the same time should have strong relatedness, and therefore share more common tags. The experiment showed that most people follow the same temporal patterns in visiting places, while the explicit and implicit features both need to be considered for tag recommendation.

3. *Friend Recommendation*
Friend recommendation analyzes the similar patterns between a target user and other users, and then recommends users with the most similar patterns to the target user. Here, the similar patterns may represent the common interests, shopping habits, traveling trajectories, etc. Friend recommendation on location-based social network mostly uses supervised learning in terms of link prediction. A set of features is firstly extracted from the historical data for each pair of users, and then a classifier is trained based on the extracted features and finally used to predict the link between two users. The social network information is used as ground truth to evaluate their proposed approaches, and ROC curves (Scellato et al. 2011c; Sadilek et al. 2012a) are usually used as evaluation metrics.

Current work on friend recommendation differs in how to choose the feature space and classifier. Chang and Sun (2011) used logistic regression to predict the link between two users who have co-locations. Feature extraction was based on the tuples of (place x, actor1, actor2), indicating that actor1 and actor2 have checked-in into place x at least once. Three features are extracted: the total number of check-ins at place x, and numbers of check-ins of actor1 and actor2 respectively. Cranshaw et al. (2010) extracted 67 features from the data on Locaccino (Sadeh et al. 2009) for each co-location observation between two users. Their features include intensity and duration, location diversity, mobility regularity, structure properties, etc., with respect to co-location properties and user attributes. Three classifiers are selected for predicting the link, while the results show that AdaBoost has the best classification performance. They also reported that there is a positive correlation between the location diversity and the number of social ties a user has in the social network, and that considering the number of co-locations between two users is not sufficient for friend recommendation.

Sadilet et al. (2012a) adopted a similar scenario, while in addition considering the content features from tweets. Scellato et al. (2011c) exploited the place features such as common check-ins, social features like common friends, and global features such as distance between homes, then adopted various classifiers in WEKA for link prediction on Gowalla. Their results demonstrated that the purely social-based features contribute least to the prediction performance, while space features and global features lead to better performance, indicating the importance of location-based activities on location-based social networking analysis.

8.3.5 Location Privacy

Location sharing is an indispensable function of location-based social networking services. Users share their locations by checking in on location-based social networking sites to let their friends know where they are and when. The location awareness can then form location-based social networks and enhance the user's social connections. For example, a user may want to hang out with his friend after learning he is nearby through his check-in status. On the other hand, while location sharing significantly enhances user experience in social networks, it also leads to privacy and security concerns. In recent years, location privacy on location-based social networks has attracted more and more attention from both academia and industry. Previous work (Lederer et al. 2003; Consolvo et al. 2005; Gundecha et al. 2011; Tsai et al. 2009) has found that privacy is a critical concern for user considering adopting location-sharing services. When using location-sharing services, some users would like to share their location with friends for social purposes, while other users may believe that sharing personal location discloses one's personal preferences and movement track, which may cause potential physical security risks. Therefore, it is inevitable to consider privacy control when designing location-sharing applications.

Researchers are interested in understanding users' preference regarding location privacy in location-based social networks, such as why people are using location-sharing services and under what circumstances they do not want to share locations, therefore improving the design of new location-sharing applications. Humphreys (2007) analyzed user behavior on Dodgeball by conducting interviews with 21 Dodgeball users, and discovered that location-based social services do influence the way people experience urban public places and their social relations. Lindqvist et al. (2011) explored how and why people use Foursquare through interviews and surveys of Foursquare users, and reported five major factors that explain the reasons: i.e., badges and fun, social connection, place discovery, keeping track of places, and competition with themselves. Furthermore, the authors also found that the majority of users had few privacy concerns, and users choose not to check in at specific locations mainly because the places are embarrassing, non-interesting, or sensitive.

Mobile applications have also been developed to help manage privacy on LBSNs. Toch et al. developed a location sharing application "Locaccino",[8] focusing on privacy

[8] http://locaccino.org

control based on the Facebook social network (Toch et al. 2010b; Sadeh et al. 2009). A Locaccino user can request the location of his Facebook friends. It allows a user to set detailed location-sharing privacy preferences, such as when and where his location can be visible to a set of pre-specified users. Toch et al. (2010a) utilized the data collected from Locaccino to investigate the location factors that influence users' location-sharing preferences. They deployed Locaccino to a set of participants, and conducted surveys on them. Their analysis showed that locations with higher location entropy (Cranshaw et al. 2010) (a measure that is utilized to evaluate the user diversity of a location: higher location entropy indicates the location has been visited by a diverse set of unique users) are more comfortable for users to share, while highly mobile users receive more requests from their friends for location sharing. Kelley et al. (2008) introduced a machine learning approach to control the sharing policy. They proposed a Gaussian Mixture based method to classify the privacy control policies of users, with evaluation on Locaccino data from 43 users and 124 pre-defined privacy policies. The prediction accuracy is chosen as the evaluation metric.

8.3.6 Related Efforts

Aside from the topics discussed in the previous sections, even more efforts have been made in mining location-based social networks. In event detection, Sakaki et al. (2010) constructed an earthquake reporting system in Japan to report earthquakes using an event detection algorithm. They considered each user who makes tweets about a target event to be a sensor of the event, and proposed a spatial–temporal model to track the event center and trajectory. De Longueville et al. (2009) utilized twitter data to analyze the spatial, temporal, and social dynamics and URL property of events related to the Marseille forest fire, aiming to investigate the potential power of leveraging Twitter for emergency planning and disaster relief.

In geographical topic analysis, researchers utilize generative models, which are combined with spatial–temporal regularities to explore the space–time structures of topical content (Pozdnoukhov and Kaiser 2011), or devised with embedded content, user preference, and geographical locations to model tweet density (Hong et al. 2012), or generated as a combination of geographical clustering and topic model to discover and compare geographical topics (Yin et al. 2011). However, among all these works, social network information is not utilized, and the evaluation of the geo-topic model is also controversial to a certain extent.

In urban computing, Cranshaw et al. (2012) developed an online system, Livehoods,[9] to explore the social dynamics of the city and reveal the different characterized regions. The authors used a spectral clustering approach to cluster the check-in locations from 18 million check-ins into different areas, with each one

[9] http://livehoods.org

representing the character of lifestyle in that area. Sadilek et al. (2012b) modeled the spread of disease through Twitter data. They proposed a detection framework to identify the sick individual based on tweet content, and showed that there is a strong correlation between a person's number of infected friends and his probability of getting sick, where the probability increases exponentially as the number of infected friends grows.

8.4 Illustrative Examples of Mining Location-Based Social Network Data

In this section, we present two examples to illustrate how to mine real-world LBSN data to improve location-based services. The first example investigates a user's social–historical ties in check-in behavior for location prediction, and the second example leverages the social network information on LBSNs to address the "cold-start" check-in problem.

8.4.1 Exploring Social–Historical Ties on Location-Based Social Networks

On location-based social networking sites, a user's check-in behavior can be analyzed as an integration of his social ties and historical ties, while both ties have varying tie strengths, as illustrated in Fig. 8.2 with the tie strengths represented by line width (Gao et al. 2010a).

1. *Discovering the Properties of Social–Historical Ties*
   The historical ties of a user's check-in behavior have two properties in LBSNs. Firstly, a user's check-in history approximately follows a power-law distribution, i.e., a user goes to a few places many times and to many places a few times. Figure 8.3a shows the distribution of check-in frequency (in log scale) on a real-world dataset[10] collected from Foursquare, with detailed dataset statistics shown in Table 8.1. The figure suggests that the check-in history follows a power-law distribution, and the corresponding exponent is approximately 1.42. The check-in distribution of an individual also shows the power-law property, as shown in Fig. 8.3b. Secondly, historical ties have a short-term effect. As illustrated in Fig. 8.2, a user arrives at the airport and then takes a shuttle to the hotel. After his dinner, he sips a cup of coffee. The historical ties of the previous check-ins at the airport, shuttle stop, hotel, and restaurant have different strengths with respect to the latest check-in at the coffee shop. Furthermore, historical tie strength decreases over time.

[10] The dataset used in this example is available at: http://www.public.asu.edu/~hgao16/dataset/SHTiesData.zip

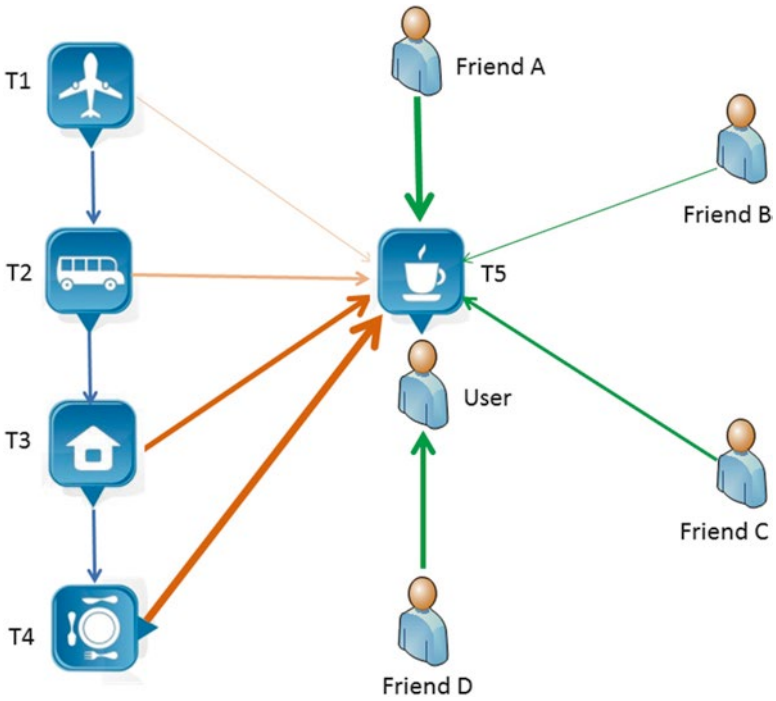

Fig. 8.2 An example: how social and historical ties may affect a user's check-ins at time T_5

To discover the properties of social ties, we compare the check-in similarity between users with friendship and those without. For each user, let $\mathbf{f} \in R^m$ be his check-in vector with the k-th element $\mathbf{f}(k)$ being the number of check-ins at location $l_k \in \pounds$, where $m = |\pounds|$ is the vocabulary size. The cosine similarity of two users u_i and u_j is defined as:

$$sim\left(u_i, u_j\right) = \frac{\mathbf{f}_i \mathbf{f}_j}{\left|\mathbf{f}_i\right|_2 \times \left|\mathbf{f}_j\right|_2}, \tag{8.2}$$

where $|\bullet|_2$ is the 2-norm of a vector.

We define the check-in similarity between u_i and a group G of other users as the average similarity between user u_i and the users in group G,

$$S_G\left(u_i\right) = \frac{\sum_{u_j \in G} sim\left(u_i, u_j\right)}{|G|}. \tag{8.3}$$

For each u_i, we calculate two similarities; i.e., $\mathbf{S_F}(u_i)$ is the average similarity of u_i and his friendship network; $\mathbf{S_R}(u_i)$ is the average similarity of u_i and a group of randomly chosen users, who are not in the friendship network of u_i. The number of randomly chosen users is the same as the amount of u_i's friends.

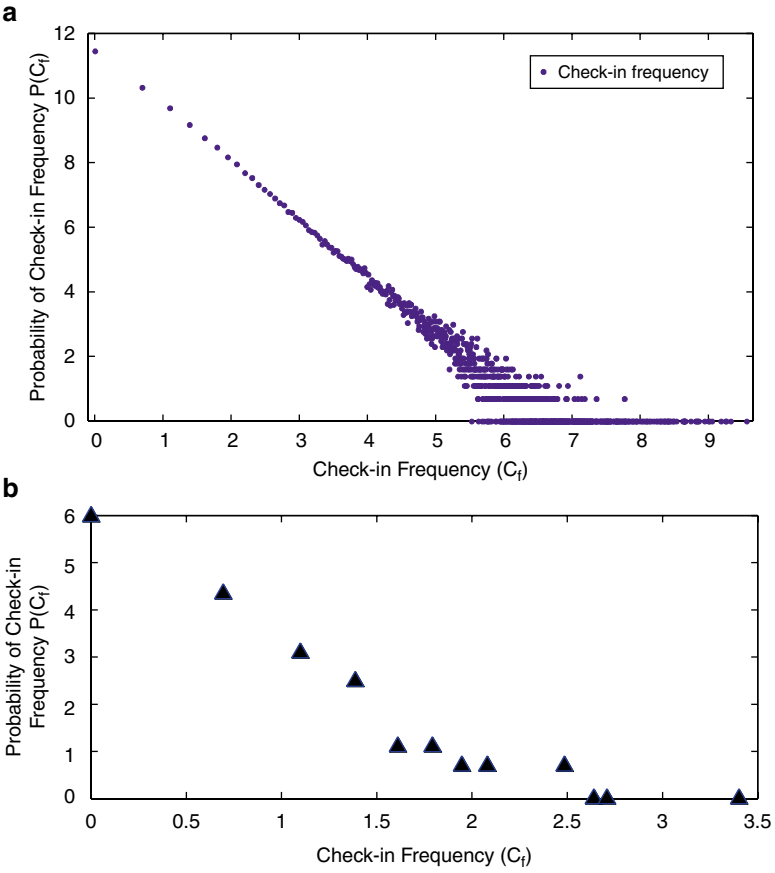

Fig. 8.3 The power-law distribution of check-ins. (**a**) Power-law distribution of check-ins in whole dataset. (**b**) Power-law distribution of check-ins in whole dataset

Table 8.1 Statistical information of Foursquare dataset

Duration	Mar. 8, 2010–Jan. 21, 2011
Number of users	18,107
Number of check-ins	2,073,740
Number of unique locations	43,063
Number of links	115,574

We conduct a two-sample *t*-test on the vectors $\mathbf{S_F}$ and $\mathbf{S_R}$. The null hypothesis is H_0: $\mathbf{S_F} \leq \mathbf{S_R}$, i.e., users with friendship share fewer common check-ins than those without, and the alternative hypothesis is H_1: $\mathbf{S_F} > \mathbf{S_R}$. In our experiment, the null hypothesis is rejected at significant level $\alpha = 0.001$ with *p*-value of 2.6*e*-6, i.e., users with friendship have higher check-in similarity than those without.

Table 8.2 Corresponding features between language and LBSN modeling

Language modeling		LBSN modeling	
Corpus		Check-in collection	
Document		Individual check-ins	
Document structure	Paragraph	Check-in structure	Monthly check-in sequence
	Sentence		Weekly check-in sequence
	Phrase		Daily check-in sequence
	Word		Check-in location

2. *Modeling Social–Historical Ties for Location Prediction*
To capture the two properties of historical ties, i.e., power-law distribution and short-term effect, a language model is utilized to model the check-in behavior. There are many features shared between language processing and LBSN mining. First, the text data and check-in data have similar structures, as shown in Table 8.2. For example, a document in language processing can correspond to an individual check-in sequence in LBSNs, while a word in the sentence corresponds to a check-in location. Second, the power-law distribution and short-term effect observed in LBSNs have also been found in natural language processing, where the word distribution is closely approximated by power-law (Zipf 1932), and the current word is more relevant to its adjacent words than distant ones. Therefore, to model the historical ties of a user, we introduce the hierarchical Pitman–Yor (HPY) language model (Teh 2006a, b) to the location-based social networks, which is a state-of-the-art language model that generates a power-law distribution of word tokens (Goldwater et al. 2006) while considering the short-term effect. We define the historical model (HM) as below,

$$P_H^i\left(c_t = l\right) = P_{HPY}^i\left(c_t = l \mid \Omega_i, \Theta\right), \tag{8.4}$$

where $P_{HPY}^i(c_t = l|\Omega_i, \Theta)$ is the probability of user u_i's check-in c_t at location l generated by the HPY with u_i's observed check-in history Ω_i, and Θ is the parameter set for the HPY language model. More technical details can be found in Gao et al. (2012a).

To model the social ties of check-in behavior, we define the social model (SM) as below,

$$P_S^i\left(c_t = l\right) = \sum_{u_j \in F(u_i)} sim\left(u_i, u_j\right) P_{HPY}^i\left(c_t = l \mid \Omega_j, \Theta\right), \tag{8.5}$$

where $F(u_i)$ is the set of u_i's friends. $P_{HPY}^i(c_t = l|\Omega_j, \Theta)$ is the probability of u_i's next check-in c_t at location l computed by HPY with u_j's check-in history Ω_j as training data. Note that only the check-ins before the prediction time are included in the training data.

Finally, a social–historical model (SHM) is proposed to explore a user's check-in behavior, integrating both historical and social effects,

$$P_{SH}^i\left(c_t = l\right) = \eta P_H^i\left(c_t = l\right) + \left(1 - \eta\right) P_S^i\left(c_t = l\right). \tag{8.6}$$

where η controls the weight from historical ties and social ties.

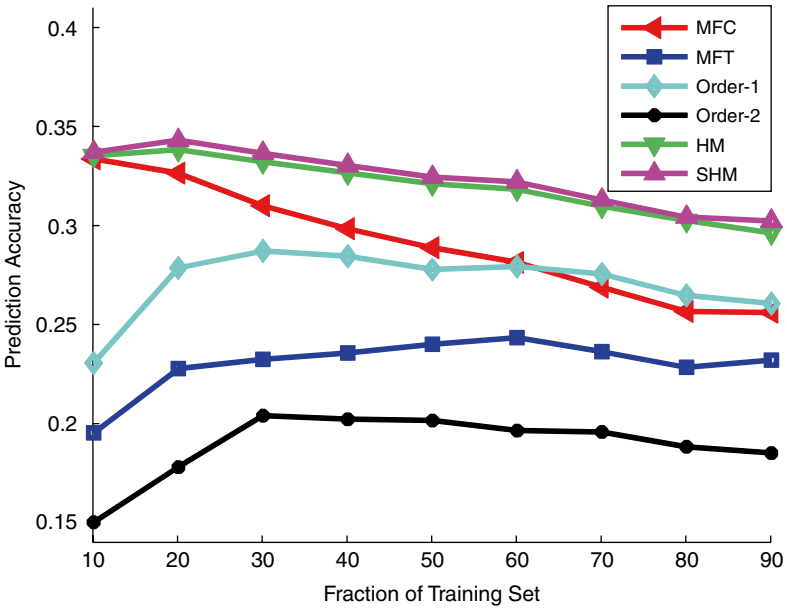

Fig. 8.4 The performance comparison of prediction models

The experimental results of location prediction on a real-world LBSN dataset are plotted in Fig. 8.4, with the performance comparison of the proposed model (HM and SHM) and four baseline models (Gao et al. 2012a). The results demonstrate that the proposed approach properly captures a user's check-in behavior by considering social–historical ties, and outperforms the current state-of-the-art prediction models.

8.4.2 gSCorr: Modeling Geo-Social Correlations for New Check-ins on Location-Based Social Networks

On location-based social networking sites, users explore various POIs and check in at places that interest them. The power-law property of users' check-in behavior in Fig. 8.3 indicates that users do visit new places, resulting in the "cold-start" check-in problem (Gao et al. 2012b). Predicting the "cold-start" check-in locations (i.e., predicting a user's next location where he has never been before) exacerbates the already difficult problem of location prediction, as there is no historical information on the user for the new place; hence, traditional prediction models relying on the observation of historical check-ins would fail to predict the "cold-start" check-ins. In this scenario, social network information could be utilized to help address the "cold-start" problem, since social theories (e.g., social correlation (Anagnostopoulos et al. 2008)) suggest that the movement of humans is usually affected by their social networks.

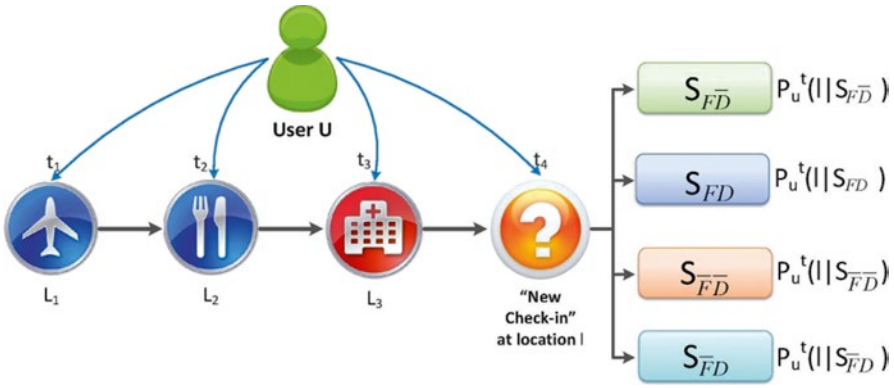

Fig. 8.5 Geo-social correlations of new check-in behavior

Table 8.3 Geo-social correlations

	F	$\bar{F}$
$\bar{D}$	$S_{F\bar{D}}$: local friends	$S_{\bar{F}\bar{D}}$: local non-friends
D	S_{FD}: distant friends	$S_{\bar{F}D}$: distant non-friends

Figure 8.5 illustrates a user's "new check-in" behavior in different social correlation aspects. User u goes to the airport at t_1, and then the restaurant at t_2 followed by the hospital at t_3. When u performs a "new check-in" at t_4, i.e., the check-in location does not belong to $\{L_1, L_2, L_3\}$, it may be correlated to those users that are from u's different geo-social circles $S_{F\bar{D}}$, S_{FD}, $S_{\bar{F}\bar{D}}$, and $S_{\bar{F}D}$, as defined in Table 8.3. Investigating these four circles enables us to study a user's check-in behavior in four corresponding aspects: local social correlation, distant social correlation, confounding, and unknown effect.

1. *Modeling Geo-Social Correlations*
   To model the geo-social correlations of "new check-in" behavior, we consider the probability of a user u checking-in at a new location l at time t as $P_u^t(l)$. We define this probability as a combination of the four geo-social correlations,

$$P_u^t(l) = \Phi_1 P_u^t\left(l|S_{F\bar{D}}\right) + \Phi_1 P_u^t\left(l|S_{F\bar{D}}\right) + \Phi_3 P_u^t\left(l|S_{FD}\right) + \Phi_4 P_u^t\left(l|S_{\bar{F}D}\right) \tag{8.7}$$

   where Φ_1, Φ_2, Φ_3, and Φ_4 are four distributions that govern the strength of different geo-social correlations, $P_u^t(l|S_x)$ indicates the probability of user u checking-in at location l that is correlated to u's geo-social circle S_x.

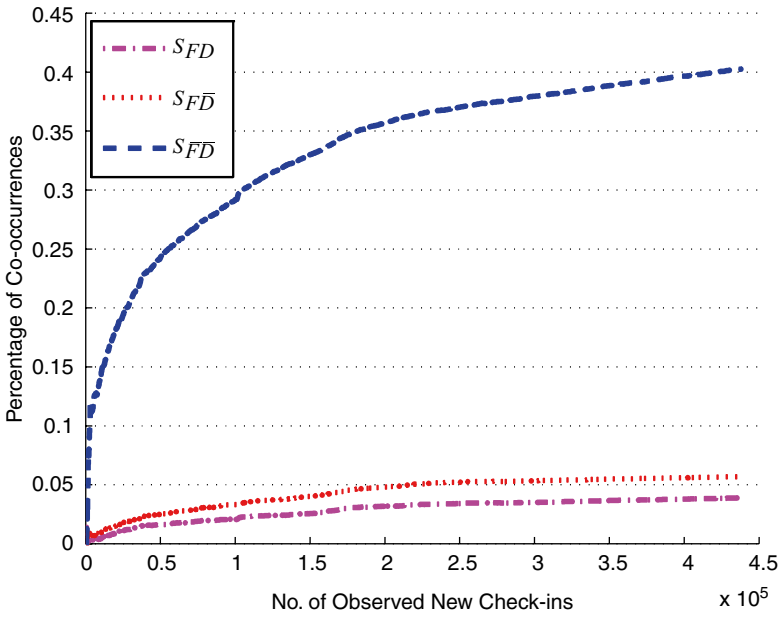

Fig. 8.6 Observed social correlations on new check-ins

Table 8.4 Statistical information of Foursquare dataset

Duration	Jan. 1, 2011–July 31, 2011
Number of users	11,326
Number of check-ins	1,385,223
Number of unique locations	182,968
Number of links	47,164

The modeling of Φ_1, Φ_2, Φ_3 and Φ_4 is based on the observation of "new check-in" distribution in Fig. 8.6, with the corresponding dataset[11] collected from Foursquare shown in Table 8.4. From Fig. 8.6, it is observed that Φ_1 is a real-valued and differentiable increasing function, and Φ_2 and Φ_3 are fairly constant. The percentage of "new check-ins" from $S_{\bar{F}D}$ is not presented, since it can be deduced from the other three. Therefore,

$$\begin{aligned}
\Phi_1 &= f\left(\mathbf{w}^T \mathbf{f}_u^t + b\right) \\
\Phi_2 &= \left(1-\Phi_1\right)\varnothing_1 \\
\Phi_3 &= \left(1-\Phi_1\right)\left(1-\varnothing_1\right)\varnothing_2 \\
\Phi_4 &= \left(1-\Phi_1\right)\left(1-\varnothing_1\right)\left(1-\varnothing_2\right) \\
0 \le \Phi_1 \le 1, 0 \le \varnothing_1 \le 1, 0 \le \Phi_1 &\le 1
\end{aligned} \tag{8.8}$$

[11] The dataset used in this example is available at: http://www.public.asu.edu/~hgao16/dataset/gScorrData.zip

Table 8.5 Check-in and social features

Features	Description
N^c	Number of check-ins in u's history
N^{nc}	Number of new check-ins in u's history
$N_{F\bar{D}}$	Number of friends in $S_{F\bar{D}}$
$N^c_{F\bar{D}}$	Number of check-ins from $S_{F\bar{D}}$
$N^{uc}_{F\bar{D}}$	Number of unique check-ins from $S_{F\bar{D}}$
$N^{vc}_{F\bar{D}}$	Number of visited check-ins from $S_{F\bar{D}}$
$N^{nvc}_{F\bar{D}}$	Number of visited unique check-ins from $S_{F\bar{D}}$
N_{FD}	Number of friends in S_{FD}
$N_{FD}{}^c$	Number of check-ins from S_{FD}
$N_{FD}{}^{uc}$	Number of unique check-ins from S_{FD}
$N_{FD}{}^{vc}$	Number of visited check-ins from S_{FD}
$N_{FD}{}^{nvc}$	Number of visited unique check-ins from S_{FD}
$N_{\bar{F}\bar{D}}$	Number of users in $S_{\bar{F}\bar{D}}$
$N^c_{\bar{F}\bar{D}}$	Number of check-ins from $S_{\bar{F}\bar{D}}$
$N^{uc}_{\bar{F}\bar{D}}$	Number of unique check-ins from $S_{\bar{F}\bar{D}}$
$N^{vc}_{\bar{F}\bar{D}}$	Number of visited check-ins from $S_{\bar{F}\bar{D}}$
$N^{nvc}_{\bar{F}\bar{D}}$	Number of visited unique check-ins from $S_{\bar{F}\bar{D}}$

where $\mathbf{f}_u{}^t$ is a check-in feature vector of a single user u at time t, $\mathbf{w}$ is a vector of the weights of $\mathbf{f}_u{}^t$, and b controls the bias. In this work, we define a user's check-in and social features $\mathbf{f}_u{}^t$ in Table 8.5. Φ_1 and Φ_2 are two constants.

To capture the geo-social correlation probabilities $P_u{}^t(l|S_x)$, three geo-social correlation measures are proposed considering the factors of location frequency, user frequency and user similarity, as described below,

- *Sim-Location Frequency (S.Lf)*

$$P_u^t\left(l \mid S_x\right) = \frac{\sum_{v \in S_x} s\left(u,v\right) N_v^t\left(l\right)}{\sum_{v \in S_x} s\left(u,v\right) N_v^t} \tag{8.9}$$

  where $s(u,v)$ represents the user similarity between user u and user v. $N_v{}^t(l)$ represents the number of check-ins at location l by user v before time t, and $N_v{}^t$ the total number of locations visited by user v that user u has not visited before time t .

- *Sim-User Frequency (S.Uf)*

$$P_u^t\left(l \mid S_x\right) = \frac{\sum_{v \in S_x} \delta_v^t\left(l\right) s\left(u,v\right)}{\sum_{v \in S_x} s\left(u,v\right)} \tag{8.10}$$

  where $\delta_v{}^t(l)$ equals to 1 if user v has checked in at l before t, and 0 otherwise.

- *Sim-Location Frequency & User Frequency (S.Lf.Uf)*

$$P_u^t\left(l \mid S_x\right) = \frac{\sum_{v \in S_x} s\left(u,v\right) N_v^t\left(l\right)}{\sum_{v \in S_x} s\left(u,v\right) N_v^t} \frac{\sum_{v \in S_x} \delta_v^t\left(l\right)}{N_{S_x}} \tag{8.11}$$

Table 8.6 Evaluation metrics

	Single measure	Various measures
Equal strength	EsSm	EsVm
Random strength	RsSm	RsVm
Various strength	VsSm	gSCorr

We adopt *S.Lf.Uf*, *S.Lf*, and *S.Uf* to compute $P_u^t\left(l \mid S_{F\bar{D}}\right)$, $P_u^t(l|S_{FD})$ and $P_u^t\left(l \mid S_{\bar{F}\bar{D}}\right)$ respectively, based on our observation of their good performance on corresponding geo-social circles. To reduce time complexity, we consider $P_u^t\left(l \mid S_{\bar{F}D}\right)$ as a probability of random jump to a location in current location vocabulary that u has not checked in before.

2. *Evaluating gSCorr*
   To evaluate gSCorr, we consider the effect of both geo-social correlation strength and measures in capturing the user's "new check-in" behavior. Therefore, we set up five baselines to compare the location prediction performance with gSCorr, as shown in Table 8.6. Each baseline adopts a different combination of correlation strength and measures, where "Es", "Rs", "Vs", "Sm", "Vm" represent "equal strength" (set all geo-social correlation strengths as 1), "random strength" (randomly assign the geo-social correlation strengths), "various strength" (the same as gScorr), "single measure" (use *S.Lf.Uf* to measure the correlation probabilities for all the geo-social circles) and "various measures" (the same as gScorr) respectively. Note that gSCorr is a various strength and various metrics approach. Following the evaluation metrics of recommendation system, we use top-k accuracy as evaluation metric and set $k=1, 2, 3$ in the experiment. For each random strength approach (RsSm and RsVm), we run *30* times and report the average accuracy.

Table 8.7 shows the detailed prediction accuracy of each method for further comparison, with the best performance highlight as italics. We summarize the essential observations below:

- The geo-social correlations from different geo-social circles contribute variously to a user's check-in behavior. Both *VsSm* and *gSCorr* perform better than their equal strength versions (i.e., *EsSm* and *EsVm*) respectively, indicating that the geo-social correlations are not equally weighted.
- The randomly assigned strength approaches (*RsSm* and *RsVm*) perform the worst compared to the other approaches, where the performance of *VsSm* has a 10.50 % relative improvement over *RsSm*, and *gSCorr* has a 26.11 % relative improvement over *RsVm*, indicating that social correlation strengths do affect check-in behavior.
- The single metric approaches (*EsSm*, *RsSm*, *VsSm*) always perform worse than the various metrics approaches (*EsVm*, *RsVm*, *gSCorr*), which suggests that for different social circles, there are different suitable correlation metrics.

gSCorr performs the best among all the approaches. To demonstrate the significance of its improvement over other baseline methods, we launch a random guess approach to predict the "new check-ins". The prediction accuracy of the random guess is always below 0.005 % for top-1 prediction, and below 0.01 % for top-2 and

Table 8.7 Location prediction with various geo-social correlation strengths and measures

Methods	Top-1(%)	Top-2(%)	Top-3(%)
EsVm	17.88	24.06	27.86
EsSm	16.20	21.92	25.43
VsSm	16.49	22.28	25.92
RsSm	14.93	20.30	23.70
RsVm	15.23	20.85	24.50
gSCorr	*19.21*	*25.19*	*28.69*

top-3 prediction, indicating that gSCorr significantly improves the baseline methods, suggesting the advantage of gSCorr as considering different geo-social correlation strength and metrics for each geo-social circle.

8.5 Conclusions and Future Work

Location-based social networks carry user-driven geographical information, and bridge the gap between real world and online social media. Typical location-based social networking sites contain a triple-layer data structure including geographical, social, and content information, providing an unprecedented opportunity for studying mobile user behavior from a spatial, temporal, and social standpoint. In this chapter, we discuss the distinct properties of location-based social network data and their challenges, and elaborate current work for data analysis and research issues on location-based social networks.

This chapter has only discussed some essential issues. There are a number of interesting directions for further exploration.

- How do we better utilize social network information on LBSNs?
  Current work (Gao et al. 2012a; Cho et al. 2011; Ye et al. 2011c) on LBSNs reports very limited contributions from social networks. In their approaches for location prediction and recommender systems, models with social network information perform slightly better than those without social information. This leads to the question "is social network information really useful in explaining human mobile behavior?". The answer is probably still "yes", but the consequent problem is how to appropriately and efficiently make use of social information in LBSNs. For example, social information could be helpful on certain specific problems, such as the "cold-start" problem (Huang et al. 2004).
- How do we handle the check-in sparseness of LBSNs?
  The sparseness of user-driven check-ins in geographical sequence in LBSNs presents challenges to application of traditional approaches that cannot handle data sparseness. For example, in Cho et al. (2011), the authors evaluate their location prediction approaches on two location-based social network datasets and one cell phone dataset, reporting significantly higher accuracy on cell phone

data compared with LBSN data. The sparseness of LBSNs data can be one of the reasons that explain this phenomenon. Finding an efficient way to handle this sparse data is very challenging.

- How do we efficiently make use of user-generated content on LBSNs?
  User-generated content such as comments and tips for locations reflects the interest of the user within a spatial–temporal context. Current work mostly focuses on geographical patterns and social contexts; very few attempts have been made to make use of the user-generated content for understanding human behavior in LBSNs. Traditional text analysis approaches in social media could be leveraged for mining such content. For example, semantic knowledge that are used to enrich short texts (Hu et al. 2009, 2011) can be utilized to analyze the tips on LBSNs. Furthermore, an interesting research direction would consider the spatial–temporal, social, and content information together for improving location-based services. Investigating such information could help design new applications more closely to a user's daily life, and therefore improve the urban experience of citizen life.

Acknowledgments This work is supported, in part, by ONR (N000141010091). The authors would like to acknowledge all of the researchers in Arizona State University's Data Mining and Machine Learning Laboratory. The views expressed in this chapter are solely attributed to the authors, and do not represent the opinions or policies of any of the funding agencies.

References

ABI Research. (2008). *Location-based mobile social networking: Hype or reality?* http://www.abiresearch.com/research/product/1002345-location-based-mobile-social-networking-hy/ Accessed 17 Nov 2012.

ABI Research. (2010). *Location-based marketing*. http://www.abiresearch.com/research/product/1005770-location-based-marketing/ Accessed 17 Nov 2012.

Anagnostopoulos, A., Kumar, R., & Mahdian, M. (2008). Influence and correlation in social networks. Las Vegas: In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 7–15).

Backstrom, L., Sun, E., & Marlow, C. (2010). Find me if you can: Improving geographical prediction with social and spatial proximity. North Carolina: In *Proceedings of the 19th International Conference on World Wide Web* (pp. 61–70).

Barnes, S., & Scornavacca, E. (2004). Mobile marketing: The role of permission and acceptance. *International Journal of Mobile Communications, 2*(2), 128–139.

Barwise, P., & Strong, C. (2002). Permission-based mobile advertising. *Journal of Interactive Marketing, 16*(1), 14–24.

Bauer, H., Barnes, S., Reichardt, T., & Neumann, M. (2005). Driving consumer acceptance of mobile marketing: A theoretical framework and empirical study. *Journal of Electronic Commerce Research, 6*(3), 181–192.

Ben-Akiva, M., Bierlaire, M., Koutsopoulos, H., & Mishalani, R. (1998). Dynamit: A simulation-based system for traffic prediction. In *DACCORS Short Term Forecasting Workshop*. The Netherlands: Citeseer.

Berjani, B., & Strufe, T. (2011). A recommendation system for spots in location-based online social networks. Salzburg: In *Proceedings of the 4th Workshop on Social Network Systems* (pp. 1–6).

Brown, C., Nicosia, V., Scellato, S., Noulas, A., & Mascolo, C. (2012). Where online friends meet: Social communities in location-based networks. Dublin: In *Sixth International AAAI Conference on Weblogs and Social Media*.

Cairncross, F. (2001). *The death of distance: How the communications revolution is changing our lives*. Boston: Harvard Business Press.

Chang, J., & Sun, E. (2011). Location 3: How users share and respond to location-based data on social networking sites. Barcelona: In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.*

Chen, G., & Kotz, D. (2000). *A survey of context-aware mobile computing research* (Technical Report No. TR2000-381). Hanover: Department of Computer Science, Dartmouth College.

Cheng, Z., Caverlee, J., & Lee, K. (2010). You are where you tweet: A content-based approach to geolocating Twitter users. Toronto: In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (pp. 759–768).

Cheng, Z., Caverlee, J., Lee, K., & Sui, D. (2011). Exploring millions of footprints in location sharing services. Barcelona: In *Proceedings of the Fifth International Conference on Weblogs and Social Media.*

Cho, E., Myers, S., & Leskovec, J. (2011). Friendship and mobility: User movement in location-based social networks. San Diego: In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1082–1090).

Consolvo, S., Smith, I., Matthews, T., LaMarca, A., Tabert, J., & Powledge, P. (2005). Location disclosure to social relations: Why, when, & what people want to share. Portland: In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 81–90).

Cranshaw, J., Toch, E., Hong, J., Kittur, A., & Sadeh, N. (2010). Bridging the gap between physical location and online social networks. Copenhagen: In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing* (pp. 119–128).

Cranshaw, J., Schwartz, R., Hong, J., & Sadeh, N. (2012). The Livehoods project: Utilizing social media to understand the dynamics of a city. Dublin: In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media: Vol. 12.*

De Longueville, B., Smith, R., & Luraschi, G. (2009). OMG, from here, I can see the flames!: A use case of mining location based social networks to acquire spatio-temporal data on forest fires. In *Proceedings of the 2009 International Workshop on Location Based Social Networks* (pp. 73–80).

Dia, H. (2001). An object-oriented neural network approach to short-term traffic forecasting. *European Journal of Operational Research, 131*(2), 253–261.

Eagle, N., Pentland, A., & Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences, 106*(36), 15274–15278.

Gao, H., Barbier, G., & Goolsby, R. (2011a). Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems, 26*(3), 10–14.

Gao, H., Wang, X., Barbier, G., & Liu, H. (2011b). Promoting coordination for disaster relief – from crowdsourcing to coordination. *Social Computing, Behavioral-Cultural Modeling and Prediction* (pp. 197–204).

Gao, H., Tang, J., & Liu, H. (2012a). Exploring social–historical ties on location-based social networks. Dublin: In *Proceedings of the Sixth International Conference on Weblogs and Social Media.*

Gao, H., Tang, J., & Liu, H. (2012b). gSCorr: Modeling geo-social correlations for new check-ins on location-based social networks. Hawaii: In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management.*

Gao, H., Tang, J., & Liu, H. (2012c). Mobile location prediction in spatio-temporal context. In *Proceedings of the Mobile Data Challenge by Nokia Workshop in conjunction with International Conference on Pervasive Computing*. Newcastle.

Gastner, M., & Newman, M. (2006). The spatial structure of networks. *The European Physical Journal B-Condensed Matter and Complex Systems, 49*(2), 247–252.

Goldenberg, J., & Levy, M. (2009). *Distance is not dead: Social interaction and geographical distance in the internet era*. Arxiv preprint arXiv:0906.3202.

Goldwater, S., Griffiths, T., & Johnson, M. (2006). Interpolating between types and tokens by estimating power-law generators. *Advances in Neural Information Processing Systems, 18*, 459.

Goodchild, M., & Glennon, J. (2010). Crowdsourcing geographic information for disaster response: A research frontier. *International Journal of Digital Earth, 3*(3), 231–241.

Gundecha, P., Barbier, G., & Liu, H. (2011). Exploiting vulnerability to secure user privacy on a social networking site. San Diego: In *Proceedings of the 17th ACM SIGKDD Conference* (pp. 511–519).

Hecht, B., Hong, L., Suh, B., & Chi, E. (2011). Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles. Vancouver: In *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems* (pp. 237–246).

Hong, L., Ahmed, A., Gurumurthy, S., Smola, A., & Tsioutsiouliklis, K. (2012). Discovering geographical topics in the Twitter stream. Beijing: In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Hu, X., & Liu, H. (2012). Text analytics in social media. In C. C. Aggawal & C. Zhai (Eds.), *Mining text data* (pp. 385–414). New York: Springer.

Hu, X., Sun, N., Zhang, C., & Chua, T. (2009). Exploiting internal and external semantics for the clustering of short texts using world knowledge. Hong Kong: In *Proceeding of the 18th ACM Conference on Information and Knowledge Management* (pp. 919–928).

Hu, X., Tang, L., & Liu, H. (2011). Enhancing accessibility of microblogging messages using semantic knowledge. Glasgow: In *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 2465–2468).

Huang, Z., Chen, H., & Zeng, D. (2004). Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. *ACM Transactions on Information Systems (TOIS), 22*(1), 116–142.

Humphreys, L. (2007). Mobile social networks and social practice: A case study of Dodgeball. *Journal of Computer-Mediated Communication, 13*(1), 341–360.

Kelley, P., Hankes Drielsma, P., Sadeh, N., & Cranor, L. (2008). User-controllable learning of security and privacy policies. Alexandria: In *Proceedings of the 1st ACM Workshop on Workshop on AISec* (pp. 11–18).

Kessler, S. (2012). *Foursquare tops 20 million users*. http://mashable.com/2012/04/16/foursquare-20-million/. Accessed 16 Nov 2012.

Kulshrestha, J., Kooti, F., Nikravesh, A., & Gummadi, K. (2012). Geographic dissection of the Twitter network. Dublin: In *AAAI International Conference on Weblogs and Social Media.*

Lederer, S., Mankoff, J., & Dey, A. (2003). Who wants to know what when? Privacy preference determinants in ubiquitous computing. Florida: In *CHI'03 Extended Abstracts on Human Factors in Computing Systems* (pp. 724–725).

Li, N., & Chen, G. (2009). Analysis of a location-based social network. Vancouver: In *International Conference on Computational Science and Engineering*: *Vol. 4*. (pp. 263–270).

Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., & Ma, W. (2008). Mining user similarity based on location history. Irvine: In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems: Vol. 34*.

Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., & Tomkins, A. (2005). Geographic routing in social networks. *Proceedings of the National Academy of Sciences, 102*(33), 11623–11628.

Lindqvist, J., Cranshaw, J., Wiese, J., Hong, J., & Zimmerman, J. (2011). I'm the mayor of my house: Examining why people use Foursquare – a social-driven location sharing application. In *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems* (pp. 2409–2418).

Mok, D., Wellman, B., & Carrasco, J. (2010). Does distance matter in the age of the Internet? *Urban Studies, 47*(13), 2747.

Monreale, A., Pinelli, F., Trasarti, R., & Giannotti, F. (2009). Wherenext: A location predictor on trajectory pattern mining. Paris: In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 637–646).

Noulas, A., Scellato, S., Mascolo, C., & Pontil, M. (2011a). An empirical study of geographic user activity patterns in Foursquare. Barcelona: In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*

Noulas, A., Scellato, S., Mascolo, C., & Pontil, M. (2011b). Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. Barcelona: In *Proceedings of SMW11*.

Noulas, A., Scellato, S., Lathia, N., & Mascolo, C. (2012). Mining user mobility features for next place prediction in location-based services. In *Proceedings of the 12th IEEE international conference on data mining* (pp. 1038–1043). Brussels.

Pozdnoukhov, A., & Kaiser, C. (2011). Space-time dynamics of topics in streaming text. Chicago: In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*: *Vol. 8*.

Rhee, I., Shin, M., Hong, S., Lee, K., Kim, S., & Chong, S. (2011). On the Levy-walk nature of human mobility. *IEEE/ACM Transactions on Networking (TON), 19*(3), 630–643.

Sadeh, N., Hong, J., Cranor, L., Fette, I., Kelley, P., Prabaker, M., & Rao, J. (2009). Understanding and capturing people's privacy policies in a mobile social networking application. *Personal and Ubiquitous Computing, 13*(6), 401–412.

Sadilek, A., Kautz, H., & Bigham, J. (2012a). Finding your friends and following them to where you are. Seattle: In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining* (pp. 723–732).

Sadilek, A., Kautz, H., & Silenzio, V. (2012b). Dublin: Modeling spread of disease from social interactions. In *Proceedings of Sixth AAAI International Conference on Weblogs and Social Media (ICWSM)*.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: Real-time event detection by social sensors. North Carolina: In *Proceedings of the 19th international conference on World wide web* (pp. 851–860).

Scellato, S., Mascolo, C., Musolesi, M., & Latora, V. (2010). Distance matters: Geo-social metrics for online social networks. Berkeley: In *Proceedings of the 3rd Conference on Online Social Networks* (pp. 8–8). USENIX Association.

Scellato, S., Musolesi, M., Mascolo, C., Latora, V., & Campbell, A. (2011a). Nextplace: A spatio-temporal prediction framework for pervasive systems. San Francisco: *Pervasive Computing*, 152–169.

Scellato, S., Noulas, A., Lambiotte, R., & Mascolo, C. (2011b). Socio-spatial properties of online location-based social networks. Barcelona: In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 329–336.

Scellato, S., Noulas, A., & Mascolo, C. (2011c). Exploiting place features in link prediction on location-based social networks. San Diego: In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1046–1054).

Scharl, A., Dickinger, A., & Murphy, J. (2005). Diffusion and success factors of mobile marketing. *Electronic Commerce Research and Applications, 4*(2), 159–173.

Spaccapietra, S., Parent, C., Damiani, M., De Macedo, J., Porto, F., & Vangenot, C. (2008). A conceptual view on trajectories. *Data and Knowledge Engineering, 65*(1), 126–146.

Tang, J., Gao, H., & Liu, H. (2012a). mTrust: Discerning multi-faceted trust in a connected world. Seattle: In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining* (pp. 93–102).

Tang, J., Gao, H., Liu, H., & Sarma, A. (2012b). eTrust: Understanding trust evolution in an online world. Beijing: In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 253–261).

Teh, Y. (2006a). A Bayesian interpretation of interpolated Kneser–Ney (Technical Report TRA2/06). Singapore: National University of Singapore.

Teh, Y. (2006b). A hierarchical Bayesian language model based on Pitman–Yor processes. Sydney: In *ACL* (pp. 985–992). Association for Computational Linguistics.

Thanh, N., & Phuong, T. (2007). A Gaussian mixture model for mobile location prediction. Hanoi: In *2007 IEEE International Conference on Research, Innovation and Vision for the Future* (pp. 152–157).

Toch, E., Cranshaw, J., Drielsma, P., Tsai, J., Kelley, P., Springfield, J., Cranor, L., Hong, J., & Sadeh, N. (2010a). Empirical models of privacy in location sharing. Copenhagen: In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing* (pp. 129–138).

Toch, E., Cranshaw, J., Hankes-Drielsma, P., Springfield, J., Kelley, P., Cranor, L., Hong, J., Sadeh, N. (2010b). Locaccino: A privacy-centric location-sharing application. Copenhagen: In *Proceedings of the 12th ACM International Conference Adjunct Papers on Ubiquitous Computing* (pp. 381–382).

Tsai, J., Kelley, P., Drielsma, P., Cranor, L., Hong, J., & Sadeh, N. (2009). Who's viewed you? The impact of feedback in a mobile location-sharing application. Boston: In *Proceedings of the 27th International Conference on Human Factors in Computing Systems* (pp. 2003–2012).

Vasconcelos, M., Ricci, S., Almeida, J., Benevenuto, F., & Almeida, V., (2012). Seattle: Tips, dones and to-dos: Uncovering user profiles in Foursquare. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining* (pp. 653–662).

Wang, F., & Huang, Q. (2010). The importance of spatial–temporal issues for case-based reasoning in disaster management. Beijing: In *2010 18th International Conference on Geoinformatics* (pp. 1–5). IEEE.

Wang, D., Pedreschi, D., Song, C., Giannotti, F., & Barab'asi, A. (2011). Human mobility, social ties, and link prediction. San Diego: In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1100–1108).

Ye, M., Yin, P., & Lee, W. (2010). Location recommendation for location-based social networks. San Jose: In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 458–461).

Ye, M., Janowicz, K., M˙ulligann, C., & Lee, W. (2011a). What you are is when you are: The temporal dimension of feature types in location-based social networks. Chicago: In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 102–111).

Ye, M., Shou, D., Lee, W., Yin, P., & Janowicz, K. (2011b) On the semantic annotation of places in location-based social networks. San Diego: In Proceedings of the 17th ACM SIGKDD *International Conference on Knowledge Discovery and Data Mining* (pp. 520–528).

Ye, M., Yin, P., Lee, W., & Lee, D. (2011c). Exploiting geographical influence for collaborative point-of-interest recommendation. Beijing: In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 325–334).

Yin, Z., Cao, L., Han, J., Zhai, C., & Huang, T. (2011). Geographical topic discovery and comparison. Hyderabad: In *Proceedings of the 20th International Conference on World Wide Web* (pp. 247–256).

Zheng, Y., Zhang, L., Xie, X., & Ma, W. (2009). Mining interesting locations and travel sequences from GPS trajectories. Madrid: In *WWW* (pp. 791–800).

Zhou, D., Wang, B., Rahimi, S., & Wang, X. (2012). A study of recommending locations on location-based social network by collaborative filtering. *Advances in Artificial Intelligence*, 255–266.

Zickuhr, K. (2012). Three-quarters of smartphone owners use location-based services. *Pew Internet & American Life Project*.

Zipf, G. (1932). *Selective studies and the principle of relative frequency in language*. Cambridge: Harvard University Press.