ABSTRACT

YUAN, GUANGCHAO. Investigating Sentiment, Homophily, and Location for Understanding User Interactions in Social Media. (Under the direction of Professor Munindar Singh.)

With the rapid development of Web 2.0 technologies, smart phones, and global position system (GPS), location sharing services have become prevalent on social media. Instead of passively absorbing information, users have become active participants in social media via various user interactions.

We divide user interactions into two categories: interactions with other users and interactions with information objects. The information objects can be locations, blogs, photos, and so on. Understanding user interactions from the massive amount of available data in social media can help bridge the gap between users' online and offline activities and thereby improve the quality of recommender systems and search engines.

There are two common characteristics of user interactions. First, user interactions usually involve data sources across multiple layers: content, social, geography, and time. Second, even though a huge amount of data is generated every day, it is sparse. Therefore, how to exploit data from various layers and overcome the data sparsity is important in understanding user interactions. We propose solutions mainly from two perspectives: exploiting (1) content, (2) *homophily*—similarity between nodes breeds connections between them.

With the ultimate goal of better understanding user interactions, this dissertation makes three main contributions. The first contribution is a framework of exploiting sentiment homophily for link prediction, with theoretical modeling and empirical evaluation. This framework helps answer the question of whether applying the homophily principle to the content would improve link prediction.

The second contribution is an approach of estimating the location of where a message originated. Due to the importance of geo-tagged messages (e.g., advertising), whether we can exploit homophily and the large amount of available content to overcome the sparsity of locations is important. Evaluation results on a Twitter dataset demonstrate that (1) the prediction error could be greatly reduced by applying homophily to both the social and the geographical layers, (2) our proposed approach of relating one user's content to another user's locations is effective in reducing prediction error.

The third contribution lies in how to exploit content and geographical homophily

to improve the performance of point-of-interest (POI) recommendation. Motivated by the sparsity of the user-POI check-in matrix, we propose a context-aware framework to improve recommendation quality. The context of a POI includes (1) aspect-based sentiment extracted from reviews and , (2) environmental context created by its surrounding POIs (*neighborhoood effect*). Our work could provide insights about how to measure a user's preference for a POI by modeling aspect-based sentiment, and how to model the neighborhood effect in general.

Investigating Sentiment, Homophily, and Location for Understanding User Interactions in Social Media

by
Guangchao Yuan

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Computer Science

Raleigh, North Carolina

2016

APPROVED BY:

Christopher Healey	Emerson Murphy-Hill
Ranga Vatsavai	Munindar Singh Chair of Advisory Committee

DEDICATION

To my husband and parents.

BIOGRAPHY

Guangchao Yuan is a member of the Service-Oriented Computing Lab in the Department of Computer Science, North Carolina State University. Before joining NC State University, she obtained her Bachelor Degree in Software Engineering from Nankai University of China in 2010. She obtained her Master Degree in Computer Science from NC State University in 2013. She spent the summer of 2013 as an intern at Xerox Research Center Webster. Her research passion lies in social media analytics, machine learning, and recommender systems.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my advisor, Dr. Munindar Singh, for his guidance and patience. His invaluable advice, consistent support, and encouragement helped me through the most difficult time of my PhD life. More importantly, the work attitude and disciplines that I learned from him would benefit me all my life. I also would like to thank my committee members, Dr. Christopher Healey, Dr. Emerson Murphy-Hill, and Dr. Ranga Vatsavai, for their kindly assistance and insightful comments on my dissertation. I would like to thank Dr. Michael Huhns for his brilliant ideas and help during the initial stage of my PhD studies.

I am thankful to my colleagues at the Service-Oriented Computing Lab, NC State University. Special thanks to Pradeep K. Murukannaiah for his various kinds of supports, i.e., collecting dataset, implementing baseline models, and editing paper submissions. Importantly, enlightening discussions with him saved me many times out of struggling. I thank Dr. Chung-Wei Hang for his guidance during the initial stage of my PhD studies. I thank Dr. Zhe Zhang for helping me learn natural language processing techniques. I also thank other members including Nirav S. Ajmeri, Anup Kalia, Dr. Pankaj R. Telang, Dr. Xibin Gao, for their insightful feedback and constant help.

I am also grateful to my external collaborators. Dr. Tong Sun at Xerox Research Center Webster provided me with a summer internship opportunity, which introduced me to the social media research area. I thank Dr. Hua Liu and Dr. Changjun Wu for providing helpful suggestion and support during our collaboration with Xerox Research Center Webster.

I would like to thank the Laboratory for Analytic Sciences at NC State University and the National Science Foundation (grant 0910868) for their support during my PhD studies. In addition, I thank Dr. Mark Wilson and Christopher Allred for their helpful comments and assistance during our collaboration.

As always, the greatest debt I owe is my family. I would like to thank my parents, Junzhou Yuan and Tan Li, for their love and patience. Last but not the least, I thank my husband, Dr. Xi Ge, for having faith in me and providing various kinds of support. Without his love and encouragement, I would not have finished this dissertation.

TABLE OF CONTENTS

LIST OF TABLES . . . . . **vii**

LIST OF FIGURES . . . . . **viii**

Chapter 1 Introduction **1**
1.1 Motivation 1
1.2 Problem 2
1.2.1 Organization 3
1.3 Contributions 7
1.3.1 Link Prediction 7
1.3.2 Location Estimation 7
1.3.3 Point-of-Interest Recommendation 8

Chapter 2 Link Prediction **10**
2.1 Introduction 10
2.2 Data and Observations 12
2.2.1 Sentiment Extraction 13
2.2.2 Observations on the Data 14
2.3 Problem Definition 14
2.3.1 Candidate Generation 15
2.4 Prediction Features 16
2.4.1 Topic-Sentiment Affiliation Construction 17
2.4.2 Sentiment Features 17
2.4.3 Structural Features 19
2.4.4 Topical Features 19
2.5 Proposed Model: TSAM 20
2.6 Experimental Evaluation 23
2.6.1 Sentiment Features Evaluation 23
2.6.2 TSAM Model Evaluation 26
2.7 Related Work 28
2.8 Conclusions 30

Chapter 3 Location Estimation **31**
3.1 Introduction 31
3.2 Data, Problem, Framework 33
3.2.1 Problem and the Percimo Framework 34
3.3 Percimo: Proposed Approach 35
3.3.1 Geo-Social Community Detection 35
3.3.2 Personal-Community Interest Detection 36

3.3.3 Location Estimation . . . 39
3.4 Evaluation . . . 40
3.4.1 Evaluation Strategy . . . 40
3.4.2 Baseline Models . . . 42
3.5 Results . . . 43
3.5.1 Percimo and Baseline Models . . . 43
3.5.2 Threshold of Defining Local Users . . . 44
3.5.3 Social and Historical Effects . . . 44
3.5.4 Symmetric Prior vs. Betweenness Centrality . . . 46
3.5.5 Evaluating Percimo on State-Level Datasets . . . 46
3.6 Related Work . . . 48
3.7 Conclusions . . . 51

Chapter 4 Point-of-Interest Recommendation . . . 52
4.1 Introduction . . . 52
4.2 Data and Problem Definition . . . 55
4.2.1 Problem Definition . . . 55
4.3 Preference Modeling via Aspect-Based Sentiment . . . 56
4.3.1 Sentiment Extraction . . . 56
4.3.2 Users Profiling . . . 57
4.3.3 POIs Profiling . . . 60
4.3.4 Aspect Matching Score . . . 60
4.4 Neighborhood Effect Modeling . . . 61
4.4.1 User-POI Preference . . . 62
4.4.2 User-Neighborhood Preference . . . 63
4.5 Evaluation of Modeling Content-based Preference and Neighborhood Effect 65
4.5.1 Results of Content-based Preference Modeling . . . 68
4.5.2 Results of Neighborhood Effect Modeling . . . 69
4.6 POI Recommendation Model . . . 71
4.6.1 Recommendation by Matrix Factorization . . . 71
4.6.2 Incorporating Content-based Preference and Neighborhood Effect 72
4.6.3 Fusing Geographical Influence . . . 72
4.7 Evaluation of POI Recommendation . . . 73
4.8 Related Work . . . 73
4.9 Conclusions . . . 78

Chapter 5 Conclusions . . . 79

BIBLIOGRAPHY . . . 82

LIST OF TABLES

Table 2.1 Graph statistics. 23
Table 2.2 Learning dataset statistics. 24
Table 2.3 F_1 scores on the positive instances for different classifiers on different combination of features. 25
Table 2.4 F_1 scores on individual sentiment feature with Random Forest classifier. 26
Table 2.5 Preprocessing parameters for evaluating TSAM. 27
Table 2.6 Evaluation results for the TSAM. 27

Table 3.1 Notation used in this paper 38
Table 3.2 Statistics of the geo-social graphs 41
Table 3.3 AEDs (km) of Percimo and baseline models 43
Table 3.4 Comparing AEDs of Percimo and other baseline models on state-level sub-datasets 47

Table 4.1 Notation used in the aspect detection model 59
Table 4.2 Representative aspects with top ranked words discovered by our aspect detection model. 68
Table 4.3 RMSE scores with and without neighborhood features. 70
Table 4.4 RMSE score on each set of neighborhood features with Linear Regression Model. 71

LIST OF FIGURES

Figure 1.1 Outline of the dissertation. 4

Figure 2.1 Probability of two users in the mention graph sharing a sentiment toward the six most frequent topics. 15
Figure 2.2 Probability of two users in the mutual-follow graph sharing a sentiment toward the six most frequent topics. 16
Figure 2.3 Graphical representation of TSAM. 21

Figure 3.1 The Percimo framework. 35
Figure 3.2 Percimo's interest-detection model. 37
Figure 3.3 Percimo's AEDs for four local-social graphs. 45
Figure 3.4 Percimo's AEDs for $\mu = 0$ (social effect), $\mu = 1$ (historical effect), and learned μ (historical & social effects). 45
Figure 3.5 Percimo's AEDs for different α_1 settings. 46

Figure 4.1 A graphical representation of our aspect detection model. 58
Figure 4.2 Number of neighborhoods in terms of different neighborhood sizes, within threshold of 100 m, 200 m and 500 m. 64
Figure 4.3 Cumulative number of users whose activity distances are less than a certain distance. 67
Figure 4.4 RMSE scores with and without ASPECT-MATCHING for user-POI preference modeling. 69

Chapter 1

Introduction

Thesis Statement: Exploiting content and homophily improves the accuracy of link prediction, reduces the prediction error of location estimation and improves the performance of location recommendation.

1.1 Motivation

With the rapid development of Web 2.0 technologies, smart phones, and global position system (GPS), location sharing services have become prevalent on social media. On the one hand, content-sharing platforms, such as Twitter and Weibo, enable users to specify their locations in profiles (e.g., Los Angeles, California), or associate a message with a *geo-tag*. We define a *geo-tag* as any representation of location, e.g., city, neighborhood, or latitude-longitude (lat-lon) coordinate. On the other hand, several location-based social networks (LBSNs) have emerged and become popular, including Yelp, Foursquare, and Facebook Place. In LBSNs, users can build connections with friends, upload photos, share locations, and leave comments about the locations. Instead of passively absorbing information, users have become active participants in social media via various user interactions.

We divide user interactions into two categories: with other users and with information objects. A user interacts with other users by sending a friend request, participating in a particular community. The information objects that a user could interact with can be a location, a blog, a photo, and so on. A user could interact with these objects by visiting

a location and sharing the location online or by commenting on, liking, or reposting a post created by others.

The unprecedented access to user-produced content, social connections, and geo-spatial trails of users provides researchers with extraordinary opportunities for investigating user interactions. Regarding users as social sensors, investigating user interactions can understand the interplay between users' online and offline activities [Cranshaw et al., 2010], recommend personalized products [Ye et al., 2012], target and spread regional advertisements [Tuten, 2008], detect earthquakes [Sakaki et al., 2010], predict political elections [Diakopoulos and Shamma, 2010], and so on.

There are two common characteristics of user interactions. First, user interactions usually involve data sources across multiple layers: content layer (*what* does a user do or say?), social layer (*who* are the friends of a user?), geographical layer (*where* does a user go?), and temporal layer (*when* does a user perform an interaction?) [Yuan et al., 2013a]; a typical user interaction requires data sources from at least two out of the four layers. For example, people meet in a social event (geographical layer) tend to become online friends in Facebook (social layer); a user may go to a restaurant (geographical layer) after seeing the review about the restaurant in Foursquare (content layer), which is posted by one of her online friends (social layer).

Second, even though a huge amount of data is generated everyday, it is sparse. For example, a social network has millions of nodes, but the links between nodes are quite sparse; in Twitter, only around 26% users have specified their locations as granular as a city name [Cheng et al., 2010], and only 2% of tweets are geo-tagged [Leetaru et al., 2013].

Therefore, how to exploit data from various layers and overcome the data sparsity is important in understanding user interactions.

1.2 Problem

At the macro level, our solutions come from two perspectives. First, exploiting information from the content layer. Compared to the other three layers, the content layer suffers least from the data sparsity problem; around 500 million tweets are generated per day

[Stats, 2016]. The content usually reflects a user's interests and sentiments. By applying text mining techniques, this vast amount of content can provide a rich source of context for understanding user interactions. In addition, building users' profiles from content can help address data sparsity (that is, sparsity in other sources).

Second, exploiting *homophily*—similarity between nodes breeds connections between them [McPherson et al., 2001]. Here a node could be a person, a location, or other entity. For user-user interactions, existing works derive similarities among users from various perspectives, including network topology [Liben-Nowell and Kleinberg, 2003], users' interests [Singla and Richardson, 2008], and geography [Scellato et al., 2011]. For user-location interactions, a common intuition is that locations within a shorter geographical distance tend to be more related than those far away [Tobler, 1970]. Thus, the geographical distance is important in users' location-visiting behavior. We claim that data sparsity can be alleviated by exploiting homophily.

Figure 1.1 outlines the main scope of the dissertation.

Overall, the dissertation investigates two kinds of user interactions—who to become online friends with (user-user interaction) and where to visit in the offline world (user-location interaction). The objective is to provide an understanding on the performance and effect of exploiting content and homophily on user interactions, through theoretical modeling and experimental evaluations.

In general, we ask two research questions:

1. How can we exploit content and homophily to improve our understanding of user interactions?
2. What's the performance gain from applying content and homophily over modeling user interactions?

1.2.1 Organization

Before the appearance of the location sharing services, the simplest user interactions in social media (e.g., Facebook or Twitter) included the following, (1) a user posts messages, which indicate her emotional status (sentiment) or her daily activities (aspect), (2) a user becomes friends with another user so that they could see each other's messages more conveniently.

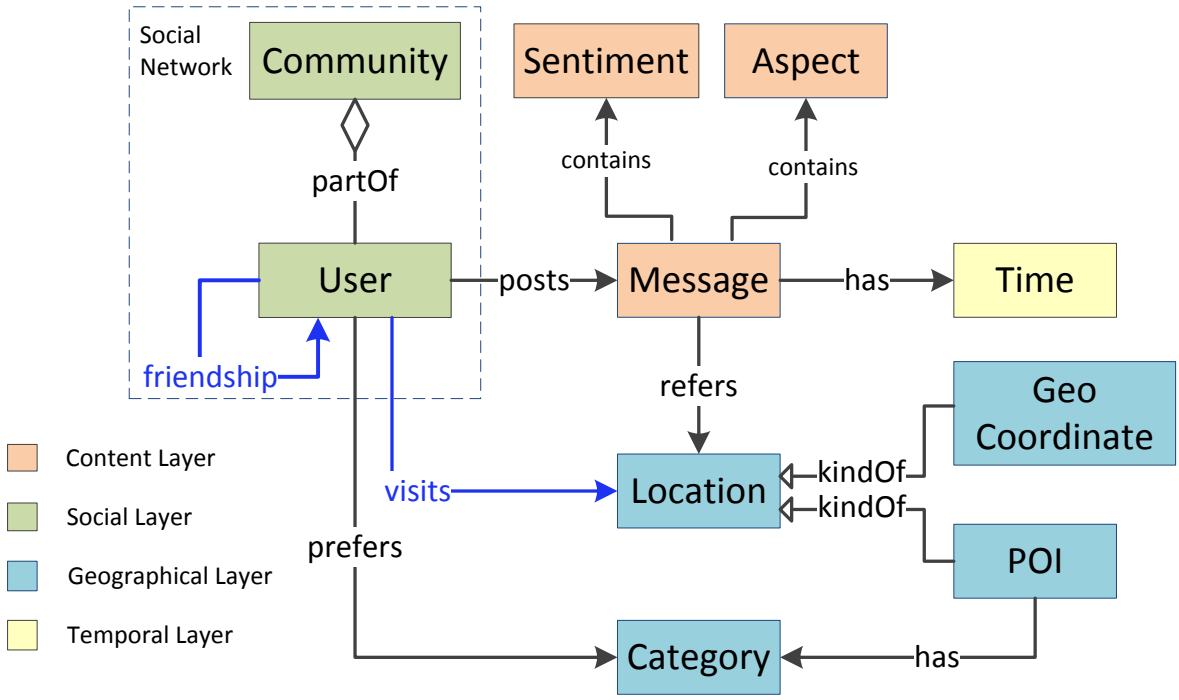

Figure 1.1 Outline of the dissertation.

An important research question that emerges from the above user interactions is *link prediction*—inferring potential relationships from a snapshot of a social network. Most existing approaches rely on deriving similarities between two users' social circles; a common intuition is that two users sharing more common friends have a higher probability of becoming friends. However, due to the difficulty of obtaining social network and the sparsity of social network, we are interested in exploiting users' messages to address the problem. A user's messages usually reveal interests and sentiment of a user, providing us with rich and unique opportunities of understanding a user. For example, during the period of political election, we can infer who a user supports for what reasons from her messages [Tumasjan et al., 2010]. Motivated by the observation that structurally or semantically similar users may express different sentiments toward a common character-

istic, we claim that there exists a *sentiment homophily* in networks—similarity in users' sentiments breeds connection. Sentiment can be an important trait for link prediction because of its effectiveness in diverse domains, such as political election prediction [Diakopoulos and Shamma, 2010], location recommendation [Yang et al., 2013], and event detection [Thelwall et al., 2011]. Therefore, our first work is to apply homophily in the content layer to improve the performance of link prediction. There are two main challenges. First, how can we quantify sentiment homophily? Second, if sentiment homophily exists, how can we leverage such homophily to build a model with the objective of improving link prediction? In Chapter 2, we discuss how to motivate the problem and address the challenges in detail. We present both the theoretical modeling and empirical evaluation of our approach on a Twitter dataset.

Next, the user interactions become complicated when the location sharing services are enabled. When a user posts a message, the message could be associated with a location. The location could be a geo-tag or a point-of-interest (POI), and the POI usually has a category (e.g., *food* or *night life*), indicating its functionality. The applications and the corresponding research questions are different depending on whether the location is a geo-tag or a POI.

If the location is a geo-tag, the geo-tag indicates where the message originated. Geo-tagged messages provide meaningful real-time information for modeling geographical phenomena, such as monitoring regional health [Aramaki et al., 2011], detecting local emergency [Starbird et al., 2010], observing linguistic differences across geographical areas [Hong et al., 2012], and so on. Nevertheless, due to the sparsity of available geo-tagged messages, *location estimation*—assigning a geo-tag to a message—has become an important research topic, which is the focus of our second work. The task of predicting the geo-tags of the large number of untagged messages, given the limited available geo-tagged messages, is challenging in two ways. First, for an untagged message of a user, how to choose a set of geo-tags that serve as the candidates? Second, how to relate a user's message to another's geo-tags? Previous approaches fall into two main categories. First, content-based techniques assume that messages encode location via place names or other location words and rely on word distributions over geo-tags. They treat historical geo-tags of all users as candidates and yield large prediction errors. Second, individualized

techniques assume that a user only visits a limited number of locations, and treat a user's prior geo-tags as candidates. The approach fails for users with insufficient historical geo-tagged tweets. Most users have sparse geo-tagged histories and some have no geo-tagged messages at all. We claim that "similar" users may have similar location visiting behavior, and a user's messages can be predicted from similar users' geo-tags. We investigate various definitions of similarity between users by applying homophily to both the social and the geographical layers, and study the effect of such definitions on the accuracy of location estimation. In addition, inspired by the existing works that a user's interests are related to the location functionality (category of a POI), we differentiate a user's personal interests from her community interests, and relate her community interests to other users' geo-tags. We present our approach and evaluation in Chapter 3.

If the location is a POI, the message is usually a user's comment about the POI. User interactions in LBSNs include checking-in at POIs, leaving their comments about POIs, and sharing them with their friends. With the popularity of LBSNs, the POI recommendation—recommending unvisited POIs to a user—has attracted attention from several researchers. POI recommender systems with improved quality not only help users explore interesting places, but also help service providers in targeting users. Most existing approaches apply matrix factorization technique [Koren et al., 2009] to user-POI check-in matrix [Ye et al., 2011; Cheng et al., 2012; Liu and Xiong, 2013; Gao et al., 2015]. However, the matrix is highly sparse since a user typically visits a few POIs, leading to poor recommendation quality. To improve the recommendation quality, we propose a framework that jointly models two contextual factors: content and geographical homophily. We observe that a user usually expresses her sentiment toward different aspects when she writes a review for a POI. For example, "Atmosphere is ok, but the pad thai here is delicious...". In the context of recommendation problem, where a user has no historical interaction with an unvisited POI, exploiting content information not only addresses the data sparsity problem to some extent, but also captures many significant qualities of a user or a POI. In addition, unlike traditional recommender systems, geographical distance brings about both challenges and opportunities. A user tends to visit POIs that are close to her home, and her preference to a POI may be affected by its nearby POIs. What are the properties of the two contextual factors and how to exploit both of them

to build the recommendation system are the main research questions. We present this work in Chapter 4.

We summarize the major results of the dissertation in Chapter 5.

1.3 Contributions

This dissertation makes the following contributions.

1.3.1 Link Prediction

Motivated by investigating *sentiment homophily* for link prediction, we are interested in two research questions:

- How may we exploit sentiment for link prediction?
- Can sentiment homophily help in link prediction?

We evaluate our approach on a dataset gathered from Twitter that consists of tweets sent in one month during U.S. 2012 political campaign along with the "follows" relationship between users. Our first contribution is defining a set of sentiment-based features that help predict the likelihood of two users becoming "friends" (i.e., mutually mentioning or following each other) based on their sentiments toward topics of mutual interest. Our evaluation in a supervised learning framework demonstrates that sentiment-based features significantly improve the performance of link prediction in terms of the F_1 measure in both mutual-follow and mention graphs. We find that the Adamic-Adar and Euclidean distance measures are the best predictors. Our second contribution is proposing a factor graph model that incorporates a sentiment-based variant of cognitive balance theory. The evaluation shows that, when tie strength is not too weak, our proposed model is more effective in link prediction than logistic regression and random forest.

1.3.2 Location Estimation

The first problem we investigate for user-location interaction is *location estimation* of tweets. We consider the following research questions:

- How to select other users that might be geographical or social related to a user?
- How to exploit the correlation between users' content and locations to estimate the location of a user's tweet? That is, *how one's interests are related to another user's locations*?

We propose an approach, *Percimo*, by (1) employing communities without exploding the set of candidate locations, (2) investigating the effect of different definitions of user similarity in location estimation, inspired from sociology, specifically, the common-bond and common-identity theory [Prentice et al., 1994], and (3) relating a user's interests to another's locations via a Latent Dirichlet Allocation (LDA) [Blei et al., 2003] based model that balances a user's personal and community interests.

We evaluate Percimo via a dataset consisting of geo-tagged tweets collected over two months from two U.S. states. We find that Percimo yields a smaller prediction error than two state-of-the-art approaches: (1) by reducing the size of candidate sets through communities, Percimo greatly reduces the prediction error compared to a purely content-based technique, (2) by differentiating a user's community interests from personal interests, Percimo reduces prediction error over baseline models relying purely on personal history, and predicts geo-tags even for users without historical geo-tags.

1.3.3 Point-of-Interest Recommendation

Next, we investigate the personalized point-of-interest (POI) recommendation problem. We consider the following research questions:

- How can we model the context in terms of content for POI recommendation?
- How can we model the context in terms of geographical homophily for POI recommendation?
- How can we exploit both of the contextual factors to improve the quality of POI recommendation?

First, we develop an LDA-based model to learn a user's category-based aspect distribution. The aspect distribution of a POI is scaled by the aspect-based sentiment of users who have visited it. We profile both users and POIs via aspect distributions, and model

a user's content-based preference toward a POI by calculating the similarity between the two distributions.

Second, we apply homophily principle to POIs. As near POIs are more related than distant POIs, we claim that a user's visiting behavior to a POI is not only decided by her preference to the POI, but also be affected by her preference for nearby POIs, which is defined as *neighborhood effect*. We discretize POIs into cells on a spatial grid with a certain threshold, and propose different sets of features to model different properties of the neighborhood effect.

Lastly, we develop a POI recommendation framework based on matrix factorization techniques by fusing the content-based preference modeling and the neighborhood effect.

We evaluate our approach on the Yelp Challenge Dataset. Our evaluation shows that (1) the aspect-based sentiment modeling could significantly improve the accuracy of POI visiting prediction, (2) the neighborhood effect arises when a user decides her POI-visiting behavior; the RMSE scores significantly decrease by exploiting neighborhood effect with the threshold 500 m of defining a neighborhood; features based on neighborhood properties (e.g., average visit) are significantly more effective in modeling the neighborhood effect than features that are based on a user's preference toward nearby POIs individually.

Chapter 2

Link Prediction

In this chapter, we study the effects of exploiting information from content layer (sentiment) and homophily principle in indicating a user-user interaction.

2.1 Introduction

Link prediction refers to inferring potential relationships from a snapshot of a social network. A common intuition behind link prediction approaches is the presence of *homophily* in networks—similarity breeds connections [McPherson et al., 2001]. Existing works derive similarities among users from network topology [Liben-Nowell and Kleinberg, 2003] (structural similarity), and users' interests [Singla and Richardson, 2008] and geography [Scellato et al., 2011] (semantic similarity).

However, structurally or semantically similar users may express different sentiments toward a common characteristic. Thelwall [Thelwall, 2010] found some evidence of both positive and negative sentiment homophily among MySpace friends. We posit that there exists a *sentiment homophily* in networks—similarity in users' sentiments breeds connection. Sentiment can be an important trait for link prediction because of its application in different domains such as political election prediction [Diakopoulos and Shamma, 2010], location recommendation [Yang et al., 2013], and event detection [Thelwall et al., 2011].

Romero et al. [Romero et al., 2013] found that topics of interest to users can predict social relationships. For example, two users interested in the topic "Obama for President"

are likely to be friends. However, the two users may exhibit the same (both support or oppose Obama) or contradictory sentiments (one supports and the other opposes Obama) toward the topic. Motivated by sentiment homophily, we imagine that the two users are more likely to become friends in the former case than in the latter. Based on the above intuition, we ask two questions:

- How may we exploit sentiments for link prediction?
- Can sentiment homophily help in link prediction?

Two challenges arise when answering the first question. First, how can we design sentiment-based features between two users in order to quantify sentiment homophily? Unlike features such as number of common friends, age, number of common places, designing sentiment features is much more complex because interpreting the sentiment of a tweet depends upon its domain and topic.

Second, employing traditional machine learning techniques (e.g., logistic regression) for link prediction assumes independence among pairs of nodes in a network, i.e., whether A–B is connected is independent of other connected pairs. However, such a case seldom exists in the real world. Heider [Heider, 1958] proposed cognitive balance theory in social psychology suggesting that if strong ties A–B and A–C exist, the likelihood of B–C becoming a tie (whether weak or strong) increases because of the "psychological strain": C will want to maintain his or her own feelings to agree with A and A's friend, B. Granovetter claimed that the B–C tie is always present in this case. The strength of a tie can be "a combination of the amount of time, the emotional intensity, the intimacy, and the reciprocal services" [Granovetter, 1973]. Therefore, we hypothesize that it is nontrivial to capture dependence between pairs of nodes (e.g., A–B and A–C ties predicting the B–C tie) via a machine learning technique. If sentiment homophily exists, can we leverage such homophily to quantify the strength of a tie? Will the sentiment-based cognitive balance theory help in link prediction? How to build such a model and how to define the strength of a tie through sentiments are our second challenge.

We employ a dataset of political tweets (and associated users) to address the second question. We extract users' sentiments toward different topics from their tweets, where sentiments are modeled as numeric scores and categorical values. Further, we design several sentiment-based features, and evaluate the effect of sentiment homophily

in a supervised setting on two social networks: the mutual-follow graph and the graph formed by users referring to each other using "@" mentions (Section 2.4). We find that sentiment-based features improve the performance of link prediction in terms of the F_1 score on both networks. We also investigate each sentiment-based feature and find that sentiment features based on the Adamic-Adar and Euclidean distance measures are the best predictors (Section 2.6.1).

We further propose a factor graph model based on Dong et al. [Dong et al., 2012], incorporating Heider's cognitive balance theory, where the strength of ties is defined based on sentiment-based features (Section 2.5). Our model outperforms the other two well-known classifiers (logistic regression and random forest) in the mutual-follow graph and in mention graphs where the strength of ties is not too weak (number of mentions exceeds three) (Section 2.6.2).

Although our analyses focus on Twitter, we conjecture that our approach can extend to a broad setting involving online information sharing, e.g., for restaurant or movie recommendations.

Contributions

Sentiment-based features. We define features that quantify the likelihood of two users becoming friends based on their sentiments toward topics of mutual interest. We evaluate the potential benefits of each feature.

Graphical model. We propose a model that incorporates the sentiment-based cognitive balance theory for link prediction. Our evaluation suggests that our model yields improved the performance (F_1 score) of link prediction when compared to traditional machine learning models.

2.2 Data and Observations

To obtain a dataset involving strong sentiments, we crawled Twitter during U.S. 2012 political campaign (from March 23 to April 23 in 2012) using the keywords "Obama" and "Romney." We preprocessed the dataset by first removing tweets that contain more than 10 hashtags. Because Twitter limits 140 characters in one tweet, a tweet containing

too many hashtags is likely to be spam [Kwak et al., 2010]. In addition, we treat users with less than five tweets as "inactive" and exclude them. The resulting dataset contains $3,970,974$ tweets from $123,073$ distinct user accounts.

Topics. A Twitter hashtag [Tsur and Rappoport, 2012] is a string beginning with "#", which is viewed as a topic marker in the tweet. Typically, users adopt the same hashtag to discuss a particular topic. Thus, we use hashtags to represent different topics.

Graphs. We investigate two kinds of undirected graphs: the mention graph and the mutual-follow graph. The mention graph is based on "@" mentions: whether a retweet, a reply, or direct reference to a user. If two users mention each other more often than a certain threshold, we create an edge in their mention graph (we experiment with multiple thresholds). In the mutual-follow graph, we create an edge between two users if they follow each other.

2.2.1 Sentiment Extraction

We use an established sentiment lexicon, SentiWordNet [Baccianella et al., 2010], to obtain the sentiment scores of all tweets. SentiWordNet contains three real-valued scores for each word in its lexicon indicating *positivity*, *negativity*, and *objectivity*; the sum of the three scores is one. In addition, we extract emoticons from tweets and estimate the three sentiment scores of each emoticon. Agarwal et al. [Agarwal et al., 2011] provide a list of emoticons and categorize them into five categories: *extremely positive*, *positive*, *neutral*, *negative*, and *extremely negative*. We assign the sentiment scores to each category as triples of positivity, objectivity, and negativity scores, respectively, $\langle 1, 0, 0\rangle$, $\langle 3/4, 0, 0\rangle$, $\langle 0, 1, 0\rangle$, $\langle 0, 0, 3/4\rangle$, and $\langle 0, 0, 1\rangle$.

We adapt the major methods described by Bakliwal [Bakliwal et al., 2013] to compute the sentiment score of a tweet. First, we choose only the adjectives in the lexicon and extract their stems [Porter, 1980] to build a pairwise stem-score mapping dictionary. We consider only adjectives because adjectives are strong indicators of sentiment [Hatzivassiloglou and Wiebe, 2000] and can improve sentiment prediction accuracy [Bakliwal et al., 2013]. Second, we use the Twitter part-of-speech tagger to extract the adjectives and emoticons in a tweet. Third, we handle the negation pattern through the Stanford parser [de Marneffe et al., 2006], which contains a dependency schema (*neg*) to indi-

cate negation. We reverse the sentiment polarity for each word marked in a *neg* schema. Finally, we obtain the positivity and negativity scores of a tweet by averaging the two scores for each adjective and emoticon; The objectivity score of the tweet is one minus the two polar scores.

2.2.2 Observations on the Data

As a sanity check on our intuition about sentiment homophily in the graphs we consider, we first determine the probability of two users sharing a same sentiment toward a topic of mutual interest, conditioned on whether they are connected.

We construct a mention graph choosing the threshold of mentions as three. That is, two users are connected if they mention each other at least three times. From the mutual-follow graph, we randomly choose a subgraph with 175 users and their friends. A pair of users is connected if there is an edge between them. For each graph, we construct pairs of unconnected users, whose number is identical to the number of connected pairs in the same graph. We choose the six most frequent topics (hashtags) in our dataset, and we want to compare the probability of two connected users sharing a sentiment with that of two unconnected users.

As Figure 2.1 shows, the probability of sharing a sentiment is 6% higher, on average, for connected users than unconnected users in the mention graph. In the mutual-follow graph (Figure 2.2), the mean difference in probability between connected and unconnected users is 4%, but the difference varies across topics. The probability difference is pronounced for "#romney" and "#santorum" (8%), whereas the difference is -1% for "#teaparty". These observations support our intuition that sentiments and connections are correlated.

2.3 Problem Definition

Let $G(V, E)$ be a social network, where V is the set of users and E is the social relationship between the users. For a given node v_s and a candidate set $C = \{v_1, v_2, \ldots, v_{|C|}\}$, our goal is to predict whether there is a link between v_s and v_i ($v_i \in C$). Specifically, the task is to find a predictive function for v_s such that: $Y = f(G, v_s, C)$, where $Y = \{y_{si} | v_i \in C\}$

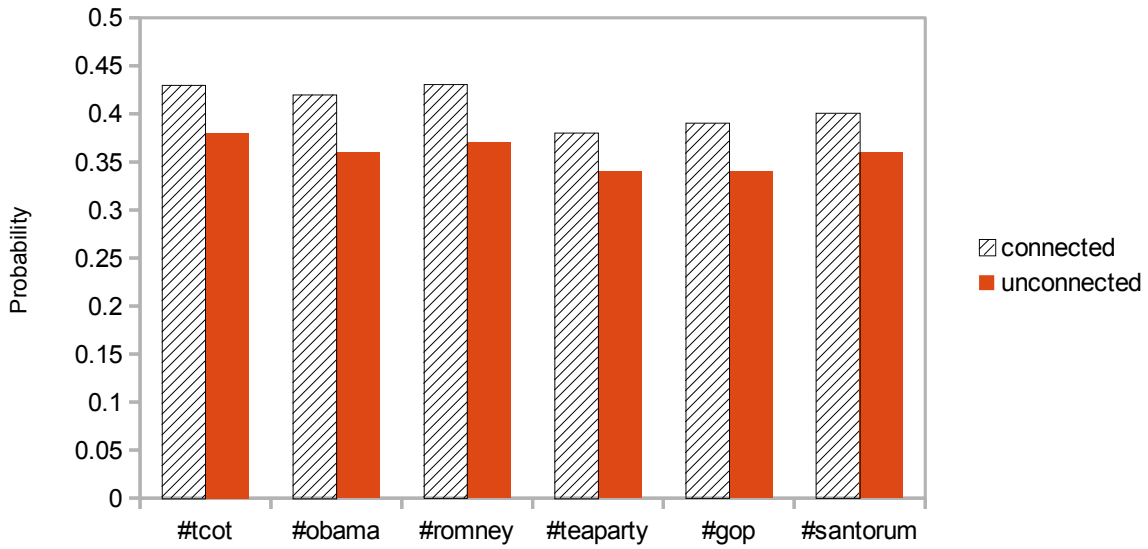

Figure 2.1 Probability of two users in the mention graph sharing a sentiment toward the six most frequent topics.

is a vector of inferred results; i.e., $y_{si} = p(1|G, v_s, C)$ represents the probability that v_s will create a link with v_i.

To do so, we take two steps. First we generate a candidate set for a source node v_s as described next. Second, we learn a predictive function by defining prediction features (Section 2.4) and applying a factor graph model incorporating cognitive balance theory (Section 2.5).

2.3.1 Candidate Generation

For each source node v_s, we choose its two-hop neighborhood as its candidate set: friends and friends of friends. We choose the two-hop neighborhood as the candidate set because (1) the number of candidates increases exponentially with the number of hops [Lichtenwalter et al., 2010]; (2) the number formed friendships decays exponentially with the number of hops [Leskovec et al., 2008].

We model our problem as a classification problem, where friends are positive instances

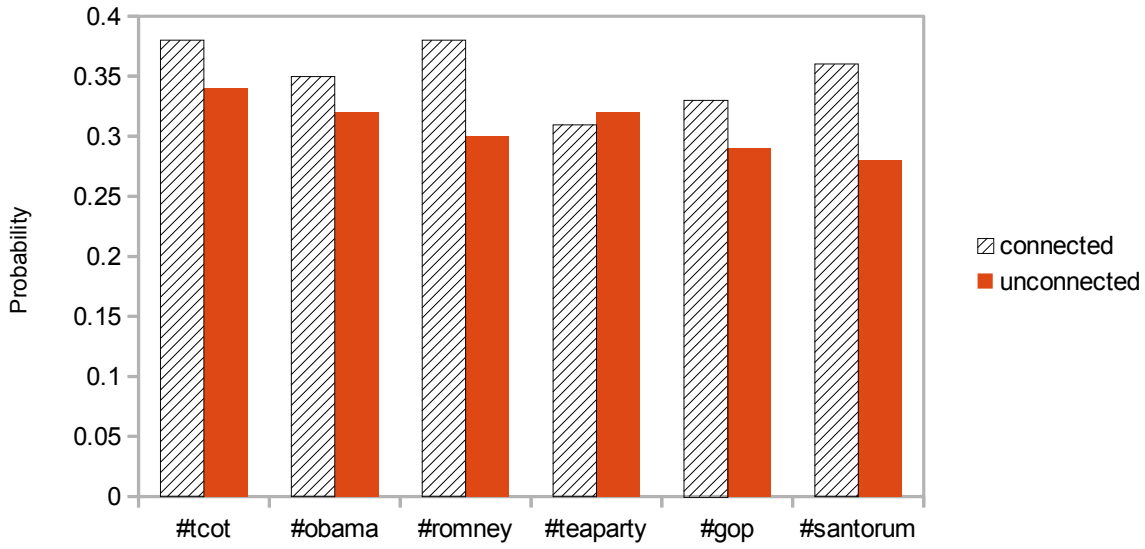

Figure 2.2 Probability of two users in the mutual-follow graph sharing a sentiment toward the six most frequent topics.

and friends of friends are negative instances. We assign half of the source nodes into a training set and half into a test set. We measure F_1 scores to validate our recommendations with respect to the existing friends of v_s in the test set.

We can derive the potential friends of v_s from its friends of friends and then make recommendations based on ranked probabilities of links between v_s and its friends of friends. However, measuring the validity of such recommendations requires that we train a model from candidates at one time instance and test for candidates at a future time instance. Since our dataset does not have information about when the links were formed, such an evaluation is out of our scope.

2.4 Prediction Features

With the development of online information sharing, the coevolution of social and affiliation networks is gaining attention, e.g., [Zheleva et al., 2009]. We consider a user as

affiliating with a topic if the user evinces interest in it, and we call the affiliation *topical affiliation*. On Twitter, topical affiliation happens when a user includes a hashtag in his or her tweet.

Combining sentiment analysis of users' messages and topical affiliation, we call such an affiliation the *topic-sentiment affiliation*. That is, a user affiliates with a set of topics, and associates a sentiment with each topic. We now describe how a user's topic-sentiment affiliation can help link prediction by describing our prediction features.

We consider three kinds of features: sentiment, structural, and topical. Sentiment features are extracted from topic-sentiment affiliation; structural features are based on the graph-based similarity between two users; topical features are based on the topical affiliation of two users, measuring the similarity in their usage of topics. The sentiment features are our contribution whereas the other two categories serve as baseline predictors.

2.4.1 Topic-Sentiment Affiliation Construction

We compute each tweet's positivity, negativity, and objectivity scores using the methods of Section 2.2.1. If a user mentions a hashtag in one of his or her tweets, we affiliate him with the topic-sentiment pair; if a user mentions the same hashtag in several tweets, we take the mean of the three scores of these tweets. We further adopt the *sentiment-volume-objectivity* (SVO) function [Gurini et al., 2013] to measure the aggregate effect of a user's level of interest and his or her sentiment scores toward a topic. The SVO score is a real value between 0 and 1 that incorporates three elements: polar sentiment (positivity or negativity), number of times a user mentions a topic, and objectivity. Therefore, the sentiment in the affiliation system consists of four numeric scores: positivity, negativity, objectivity, and the SVO score. In this way, we obtain a topic-sentiment affiliation for each user from his or her tweets.

2.4.2 Sentiment Features

We use the difference between the positivity and negativity scores to decide a user's categorical sentiment toward a hashtag. The opinion is positive (negative) if the difference is greater (less) than zero; otherwise, the opinion is objective. In addition, the *size* of a

hashtag is the number of users who have adopted it. We adopt the following notation:

- Let v_1, v_2, ..., v_N be N users.
- Let h_1, h_2, ..., h_M be M hashtags.
- Let P_i, N_i, O_i be user v_i's adopted hashtags set with positive, negative, objective sentiments, respectively.
- Let $\overline{P_j}$, $\overline{N_j}$, $\overline{O_j}$ be the set of users who expressed positive, negative, objective sentiments in a hashtag h_j, respectively.
- Let $H_i = P_i \cup N_i \cup O_i$
- Let $U(h_j) = \overline{P_j} \cup \overline{N_j} \cup \overline{O_j}$
- Let $s_i(h_j)$ be user v_i's SVO score toward hashtag h_j.

Given two users v_s and v_t, we divide sentiment features into the following seven categories:

1. **The number of hashtags for which they have the same sentiments**.
   - SENTIMENT-AGREEMENT: $|P_s \cap P_t| + |N_s \cap N_t| + |O_s \cap O_t|$
2. **Sentiment alignment coefficient**: among the common hashtags, the number that involve the same or opposite polar sentiments.
   - SENTIMENT-ALIGNED: $(|P_s \cap P_t| + |N_s \cap N_t|)/|H_s \cap H_t|$
   - SENTIMENT-MISALIGNED: $(|P_s \cap N_t| + |N_s \cap P_t|)/|H_s \cap H_t|$
3. **Size of the rarest common hashtags**: among the hashtags for which the users share polar sentiments, the one that is the least adopted by all of them.
   - SENTIMENT-RAREST: $\min(\min_{h_j \in P_s \cap P_t} |\overline{P_j}|, \min_{h_j \in N_s \cap N_t} |\overline{N_j}|)$
4. **Adamic-Adar**: sum of the Adamic-Adar distances for each hashtag set affiliated with the three opinions.
   - SENTIMENT-AA: $\sum_{h_j \in P_s \cap P_t} 1/\log |\overline{P_j}| + \sum_{h_j \in N_s \cap N_t} 1/\log |\overline{N_j}| + \sum_{h_j \in O_s \cap O_t} 1/\log |\overline{O_j}|$
5. **Sum of inverse size**.

- SENTIMENT-INVERSE: $\sum_{h_j \in P_s \cap P_t} 1/|\overline{P_j}| + \sum_{h_j \in N_s \cap N_t} 1/|\overline{N_j}| + \sum_{h_j \in O_s \cap O_t} 1/|\overline{O_j}|$

6. **Mean size of common hashtags for which they share the same sentiment**.
   - SENTIMENT-MEAN: $\frac{1}{\text{SENTIMENT-AGREEMENT}} \times (\sum_{h_j \in P_s \cap P_t} |\overline{P_j}| + \sum_{h_j \in N_s \cap N_t} |\overline{N_j}| + \sum_{h_j \in O_s \cap O_t} |\overline{O_j}|)$
7. **Topic-SVO distance**.
   - EUCLIDEAN: $\sqrt{\sum_{h_j \in H_s \cap H_t} (s_s(h_j) - s_t(h_j))^2}$
   - COSINE: $d_s d_t / \|d_s\| \|d_t\|$, where d_s and d_t are the SVO score vectors for common hashtags

2.4.3 Structural Features

These features are based on graph structure without considering semantic information. We choose four predictors introduced by Liben-Nowell et al. [Liben-Nowell and Kleinberg, 2003].

1. Number of common neighbors (CN) between two users.
2. Jaccard's coefficient (JC): CN divided by the total number of neighbors.
3. Adamic-Adar (STRUCTURAL-AA) [Adamic and Adar, 2001]: weighting the importance of a common neighbor by the degree of the neighbor.
4. Preferential attachment (PA): the product of two users' degrees.

2.4.4 Topical Features

These features, introduced by Romero et al. [Romero et al., 2013], are baseline predictors in my study.

1. The number of common hashtags (COMMON): $|H_s \cap H_t|$
2. Size of the smallest common hashtag (SMALLEST): $\min_{h_j \in H_s \cap H_t} |U(h_j)|$

3. Adamic-Adar distance (TOPICAL-AA):
   $\sum_{h_j \in H_s \cap H_t} 1/\log |U(h_j)|$
4. Sum of inverse sizes (INVERSE): $\sum_{h_j \in H_s \cap H_t} 1/|U(h_j)|$
5. Mean size of common hashtags (MEAN):
   $\frac{1}{|H_s \cap H_t|} \sum_{h_j \in H_s \cap H_t} |U(h_j)|$

2.5 Proposed Model: TSAM

To investigate how to better exploit sentiments for link prediction, we propose a topic-sentiment affiliation based graphical model (TSAM). The motivation underlying TSAM arises from cognitive balance theory: if A–B and A–C are strong ties, then these two links are not independent because B–C is likely to be present. We seek (1) a way of building such relationships where the strength of a tie incorporates sentiment and (2) to study whether this sentiment-based cognitive balance theory could improve link prediction.

A TSAM model is an undirected graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ represents the set of variables and $\mathcal{E}$ is the set of edges in the graphical model. Below, *variables* and *edges* refer to entities in TSAM, and *nodes* and *links* refer to entities in the social network.

We now describe how we build a TSAM model $\mathcal{G}(\mathcal{V}, \mathcal{E})$ by representing links and their relationships in a social network $G(V, E)$. Given the social network $G(V, E)$, a source node v_s and its candidate set $C = \{v_1, v_2, \ldots, v_{|C|}\}$, link prediction seeks to infer the probability y_{si} that v_s will create a link with v_i. Thus, we treat y_{si} as a variable (hidden) and the relationships between such variables as edges $\mathcal{E}$ in the TSAM model.

There is an edge between any two hidden variables if they contain the same source node v_s. Thus, for each source node, there is a clique in the TSAM model. For each hidden variable, an observed variable is connected with it, representing a vector of features associated with the hidden variable.

For example, suppose v_s has five candidates: $\{v_1, v_2\}$, its one-hop friends, and $\{v_3, v_4, v_5\}$, its friends of friends. The resulting TSAM model of Figure 2.3 has five hidden variables $\{y_{s1}, \cdots, y_{s5}\}$ and five observed variables $\{x_{s1}, \cdots, x_{s5}\}$.

Even though the figure only shows the factor graph for one source node, our model captures a scenario with multiple source nodes. Correspondingly, a TSAM model with

multiple source nodes is composed of multiple disconnected components, where all the hidden variables in one component form a clique.

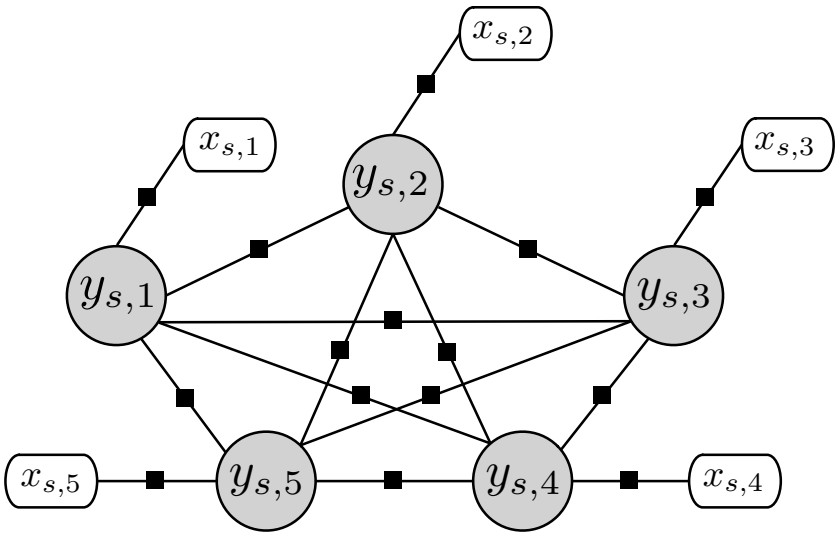

Figure 2.3 Graphical representation of TSAM.

Let T represent the set of indices for any link between a source node and one of its candidates. Then $Y = \{y_t | t \in T\}$ and $X = \{x_t | t \in T\}$ denote the set of hidden and observed variables in the TSAM model, respectively.

The graph can be modeled as a conditional random field [Lafferty et al., 2001] that defines a distribution over the graph:

$$P(Y|X) = \frac{1}{Z} \prod_{t \in T} \varphi(y_t, x_t) \prod_{(y_t, y_{t'}) \in \mathcal{E}} \psi(y_t, y_{t'})$$

where Z is a constant that ensures $\sum_Y P(Y|X) = 1$.

The model incorporates two factor functions, which we instantiate by the Hammersley-Clifford theorem [Hammersley and Clifford, 1971]. We follow the presentation of Dong et al. [Dong et al., 2012]: the **attribute factor** models the influence of different features on the hidden variable (link).

$$\varphi(y_t, x_t) = e^{\{\sum_{m=1}^{d} \alpha_m f_m(y_t, x_{t_m})\}} \tag{2.1}$$

where α_m is a weight constant and d is the number of features associated with y_t. We include all the features in Section 2.4 to build the attribute factor. Second, the **edge factor** encodes the relationships between connected hidden variables.

$$\psi(y_t, y_{t'}) = e^{\{\sum_{n=1}^{k} \beta_n g_n(y_t, y_{t'})\}} \tag{2.2}$$

where β_n is a weight constant and k is the number of features associated with the edge $(y_t, y_{t'})$ in the TSAM model.

We define $g_n(y_t, y_{t'})$ as a binary function. For any two hidden variables y_t and $y_{t'}$, a triad is potentially involved because y_t and $y_{t'}$ contain the same source node. The triad would be cognitively balanced if both y_t and $y_{t'}$ are strong ties. We can use any sentiment feature that satisfies the axioms of Gupte and Eliassi-Rad [Gupte and Eliassi-Rad, 2012] to define tie strength. Let's take the feature SENTIMENT-AA as an example. Following Hopcroft et al.'s [Hopcroft et al., 2011] definition of the importance of an user, we select the top 1% edges in the social network $G(V, E)$ in terms of SENTIMENT-AA features as strong ties. Therefore, $g_n(y_t, y_{t'})$ is one when both y_t and $y_{t'}$ are strong ties; otherwise, it is zero.

Accordingly, the log-likelihood objective function is

$$\mathcal{O}(\theta, Y|X) = \log P(Y|X) = \sum_{i \in T} \sum_{m=1}^{d} \alpha_m f_m(y_i, x_{i_m}) +$$

$$\sum_{(i,j) \in \mathcal{E}} \sum_{n=1}^{k} \beta_n g_n(y_i, y_j) - \log Z$$

Here $\theta = (\alpha, \beta)$ is the model (parameter configuration) that I seek to learn to maximize the log-likelihood objective function: $\theta^\star = \arg\max_\theta \mathcal{O}(\theta, Y|X)$.

We adapt the methods in [Tang et al., 2011] to learn the model except that we conduct experiments in a supervised setting; we use gradient descent to optimize the objective function, where the gradient is approximated by loopy belief propagation [Murphy et al., 1999]. With the estimated model $\theta^\star$, the goal of link prediction is to determine the probabilities of hidden variables that maximize the joint probability $P(Y|X, \theta^\star)$:

$$Y^\star = \arg\max_Y P(Y|X, \theta^\star) \tag{2.3}$$

2.6 Experimental Evaluation

We seek to answer two questions:

- Do sentiment features help in link prediction?
- Does our proposed graphical model incorporating the sentiment-based cognitive balance theory benefit link prediction?

2.6.1 Sentiment Features Evaluation

Evaluation Strategy. We conduct our experiments using the mention and mutual-follow graphs. For the mention graphs, the number of "@" references between users can be viewed as the strength of a tie. We therefore define several mention graphs by setting different strengths of ties. Table 2.1 shows the statistics of mention graphs with different thresholds and of the mutual-follow graph.

Table 2.1 Graph statistics.

Graph	Nodes	Edges	Mean degree
@ ≥ 1	23,915	53,009	4.43
@ ≥ 2	8,936	12,978	2.90
@ ≥ 3	4,770	5,933	2.49
@ ≥ 5	2,106	2,134	2.03
@ ≥ 7	1,201	1,100	1.83
@ ≥ 9	762	639	1.68
Follow	11,239,979	48,572,793	8.64

For each mention graph, we choose all users whose degree is less than 50 as v_s. For the mutual-follow graph, we select users who adopt at least 50 hashtags; additionally, we only select "active" users, with the degree falling within range $[50, 100]$. Then we generate the two-hop candidate set for each source node. A pair is constructed between each source node and any one of its candidates. Doing so leads to the class imbalance

being extremely high, a common problem in link prediction [Lichtenwalter et al., 2010]. Because we first want to investigate the effect of sentiment features in link prediction, we undersample the negative instances to obtain a balanced dataset. Table 2.2 shows the number of instances of all the learning datasets after preprocessing.

Table 2.2 Learning dataset statistics.

Graph	Overall	Positive	Negative
@ ≥ 1	183,359	78,301	105,058
@ ≥ 2	46,357	22,087	24,270
@ ≥ 3	21,816	10,763	11,053
@ ≥ 5	7,429	3,981	3,448
@ ≥ 7	3,653	2,094	1,559
@ ≥ 9	2,064	1,232	832
Follow	11,153	5,193	5,960

We normalize all the features to $[0, 1]$. We apply logistic regression and random forest models from the WEKA framework, and we conduct a 10-fold cross-validation with default parameters. The F_1 score and the Area under the Receiver-Operating-Characteristic Curve (AUC) are two widely used metrics in evaluating performance of classifiers [Daskalaki et al., 2006]. In our setting, we are more interested in the *true positive* than the *true negative* metric for two reasons. First, positive instances (links between source nodes and their existing friends) are ground truth, and ensuring high *recall* of positive instances is nontrivial. Second, *false positive* with zero, indicating no new friends to recommend, is not necessarily desirable. Because AUC incorporates both positive and negative instances equally whereas F_1 ignores the true negative metric, we adopt F_1 as our performance metric.

Results. Table 2.3 shows F_1 scores on different combinations of features for logistic regression and random forest classifiers. T, SE, ST represent topical, sentiment, structural features, respectively. In general, sentiment features yield better performance in terms of F_1 scores, no matter whether they are combined with structural features to build

the model. To investigate whether sentiment features indeed help improve the F_1 score, we conduct a paired t-test: each value in sample one is the F_1 score with sentiment or sentiment plus structural features; each value in sample two is the F_1 score with topical or topical plus structural features. The p-value is 0.0012, indicating the difference is statistically significant. In addition, structural features perform much better than both sentiment and topical features, but adding sentiment features can generally improve performance. Thus, sentiment features can indeed help in link prediction, but as adjuncts to the structural features.

Table 2.3 F_1 scores on the positive instances for different classifiers on different combination of features.

Classifier	Feature Sets	@ ≥ 1	@ ≥ 2	@ ≥ 3	@ ≥ 5	@ ≥ 7	@ ≥ 9	Follow
Logistic Regression	T	0.295	0.489	0.530	0.650	0.720	0.735	0.479
	SE	0.312	0.500	0.550	0.650	0.712	0.739	0.499
	ST	0.736	0.740	0.759	0.792	0.812	0.846	0.666
	T+ST	0.689	0.777	0.792	0.818	0.846	0.880	0.678
	SE+ST	0.717	0.789	0.800	0.834	0.861	0.877	0.679
	All	0.705	0.778	0.795	0.838	0.858	0.880	0.682
Random Forest	T	0.628	0.765	0.804	0.839	0.853	0.880	0.516
	SE	0.723	0.818	0.845	0.859	0.876	0.883	0.517
	ST	0.875	0.908	0.913	0.939	0.946	0.957	0.667
	T+ST	0.914	0.946	0.956	0.971	0.972	0.972	0.680
	SE+ST	0.916	0.949	0.959	0.972	0.975	0.979	0.685
	All	0.916	0.949	0.958	0.972	0.970	0.979	0.684

Individual Feature Evaluation. We show the performance of each sentiment feature for random forest since it outperforms logistic regression in Table 2.3. Table 2.4 regression in Table 2.3. Table 2.4 shows the F_1 score for individual sentiment feature. We highlight the top three ranked features in each graph. We find that in the mention graphs, the features EUCLIDEAN, SENTIMENT-MEAN, and SENTIMENT-AA perform best. We find that in the mutual-follow graph, the feature SENTIMENT-AGREEMENT performs best, but

EUCLIDEAN and SENTIMENT-AA are good indicators. And, SENTIMENT-REREST is not as informative as SMALLEST (topical features). In addition, SENTIMENT-ALIGNED outperforms SENTIMENT-MISALIGNED in general in all graphs, indicating the existence of sentiment homophily. Therefore, our evaluation suggests that sentiment homophily exists and can benefit link prediction.

Table 2.4 F_1 scores on individual sentiment feature with Random Forest classifier.

Feature	@ ≥ 1	@ ≥ 2	@ ≥ 3	@ ≥ 5	@ ≥ 7	@ ≥ 9	Follow
SENTIMENT-AGREEMENT	0.166	0.430	0.498	0.636	0.696	0.711	**0.488**
SENTIMENT-ALIGNED	0.192	0.291	0.555	0.650	0.700	0.735	0.452
SENTIMENT-MISALIGNED	0.122	0.509	0.526	0.640	0.689	0.710	0.390
SENTIMENT-REREST	0.122	0.307	0.360	0.691	0.721	0.742	0.295
SENTIMENT-AA	**0.430**	**0.589**	**0.706**	**0.753**	**0.789**	**0.822**	**0.466**
SENTIMENT-INVERSE	0.349	0.501	0.648	0.741	0.785	0.812	0.452
SENTIMENT-MEAN	**0.455**	**0.593**	**0.706**	**0.757**	**0.791**	**0.825**	0.454
EUCLIDEAN	**0.589**	**0.721**	**0.751**	**0.788**	**0.825**	**0.832**	**0.467**
COSINE	0.247	0.480	0.525	0.625	0.692	0.748	0.456

2.6.2 TSAM Model Evaluation

Following the preprocessing strategy proposed by Backstrom and Leskovec, we choose "active" source nodes in all graphs. That is, active nodes are those whose degree is within the range [lower, upper]. After constructing pairs consisting of each source node and each of its candidates, we remove those whose number of common friends is less than a threshold, because users with only a few common friends are unlikely to form friendships. Table 2.5 shows the criteria we used. The criteria differ for graphs with differing statistics. As the ties become stronger, the mention graph becomes smaller. Thus an overfitting problem may arise if the dataset is too small. We therefore limit our evaluation to graphs with more than 100 source nodes left after preprocessing.

Because the feature SENTIMENT-AA ranks in the top three features for each graph in

Table 2.5 Preprocessing parameters for evaluating TSAM.

Graph	Lower	Upper	Threshold
@ ≥ 1	10	50	4
@ ≥ 2	10	50	4
@ ≥ 3	5	50	4
@ ≥ 5	3	50	2
@ ≥ 7	3	50	2
@ ≥ 9	3	50	2
Follow	50	100	4

Table 2.6 Evaluation results for the TSAM.

Feature	Logistic Regression			Random Forest			TSAM		
	P	R	F_1	P	R	F_1	P	R	F_1
@ ≥ 1	0.350	0.443	**0.391**	0.445	0.196	0.272	0.193	0.890	0.317
@ ≥ 2	0.456	0.780	**0.576**	0.616	0.535	0.573	0.602	0.539	0.569
@ ≥ 3	0.727	0.596	0.655	0.667	0.628	0.647	0.712	0.620	**0.663**
@ ≥ 5	0.752	0.701	0.726	0.863	0.806	0.834	0.859	0.863	**0.861**
Follow	0.794	0.636	0.705	0.759	0.664	0.708	0.677	0.801	**0.734**

Table 2.4, we choose it to define the strength of a tie in the edge-factor function. For each graph, we assign half of the source nodes into a training and half into a test set. In the training phase, we set the learning rate $\lambda = 0.001$ and the number of iterations as 500. We split each graph five times and report the mean precision, recall, and F_1 scores of the TSAM model as well as for logistic regression and random forest, and identify the best performing model in terms of F_1. In Table 2.6, P and R represent precision and recall, respectively. TSAM outperforms the other two models in the mutual-follow graph. In the mention graphs, TSAM performs best when the ties are strong (@ ≥ 3 and @ ≥ 5), but not for weak ties (@ ≥ 1 and @ ≥ 2). This suggests that when the ties are weak, sentiment-based cognitive balance theory does not help because the links between users are somewhat random; hence, a balanced cognitive structure does not necessarily mean any relationship between two pairs.

Overall, our results indicate that performance of link prediction could be improved when we incorporate sentiment-based cognitive balance theory, especially on graphs where the strength of relationship is not too weak (mutual-follow graph or mention graphs where the number of mentions exceeds three).

2.7 Related Work

Link Prediction. Existing work on link prediction can be classified into two categories. For the unsupervised methods, Liben-Nowell and Kleinberg evaluated different "proximity" features extracted from network topology. The intuition is to compute the similarities on pair of nodes, and a pair with a higher similarity has a higher probability of becoming a link. They found that the Adamic-Adar metric performs best in predicting links.

Most recent works are based on supervised methods. Lichtenwalter et al. [Lichtenwalter et al., 2010] provided a detailed analysis of challenges, such as class imbalance, of using supervised methods in link prediction. Depending on the techniques, supervised methods can be further divided into three categories: feature-based classification, probabilistic graph model, and matrix factorization.

The main focus of the feature-based classification methods is to define and extract a set of appropriate features between a pair of nodes. Leskovec et al. [Leskovec et al., 2010] proposed a collection of features based on degrees of nodes and the two-step paths to predict positive or negative links on social network. A positive link indicates friendship or approval, whereas a negative link suggests disagreement or distrust. With the same objective, Chiang et al. [Chiang et al., 2011] showed that features derived from longer cycles could achieve state-of-the-art performance. Lu et al. [Lu et al., 2010] proposed a variety of path-based features derived from multiple sources and a feature selection strategy based on structural sparsity. Scellato et al. [Scellato et al., 2011] exploited place features in predicting links on location-based social networks.

Some researchers solved the link prediction problem based on probabilistic graph model. Clauset et al. [Clauset et al., 2008] proposed a probabilistic model to infer hierarchical structure from social network, and demonstrated that the hierarchical structure can be used to predict links with high accuracy, compared with state-of-the-art techniques.

Leroy et al. [Leroy et al., 2010] proposed a two-phase method based on the bootstrap probabilistic graph to address the cold start link prediction problem. Yang et al. [Yang et al., 2011] developed a joint friendship-interest propagation model that leverages the correlation between friendship and interest to address both tasks in one unified framework. Backstrom and Leskovec [Backstrom and Leskovec, 2011] developed a supervised random walk algorithm for friend recommendation on Facebook. Dong et al. [Dong et al., 2012] proposed a probabilistic graphical model to predict links.

Lastly, Menon and Elkan [Menon and Elkan, 2011] extended matrix factorization techniques to solve the link prediction problem.

Whereas we adopt a supervised approach, we additionally consider sentiment features and investigate how they improve link prediction.

Collaborative Tagging Systems. These are based on a tripartite structure: users, tags, and resources, enabling users to share their tags for particular resources. Combined with social structure, collaborative tagging systems provide new modalities of link prediction. Marlow et al. [Marlow et al., 2006] found that users tend to have a larger similarity of tag vocabularies with their friends compared with random users. Markines et al. [Markines et al., 2009] built a foundation for the folksonomy-based similarity measures, such as matching, overlap, Jaccard, and cosine similarity. Romero et al. [Romero et al., 2013] studied the relationship between topical affiliations and social network on Twitter. They found that the adoption of hashtags can predict users' social relationships. We design sentiment features based on Romero et al.'s findings. Further, we propose a graphical model based on sentiment features.

Sentiment Analysis. Sentiment analysis is an ongoing research field of text mining. With the popularity of social media, sentiment analysis brings us deeper understanding of social network analysis. Twitter enables researchers to access huge amounts of data to discover collective sentiments [Bollen et al., 2011; Thelwall et al., 2011], predict political elections [Diakopoulos and Shamma, 2010; Tumasjan et al., 2010], and so on. In other applications, Tan et al. [Tan et al., 2011] used the social relationship to improve user-level sentiment prediction. Some researchers extracted sentiments from users' reviews on POIs to improve location recommendation services [Yang et al., 2013; Gao et al., 2015; Zhang et al., 2015]. We conduct our work with a different purpose: using sentiment homophily

to predict links.

2.8 Conclusions

We study how to exploit sentiments for link prediction, and evaluate the extent to which sentiment homophily can help improve link prediction. By extracting users' sentiments from their tweets on different topics, we describe a set of sentiment features to quantify the likelihood of two users becoming friends. The evaluation results suggest that sentiment features significantly improve the performance of link prediction in terms of F_1 in both mutual-follow and mention graphs. We find that Adamic-Adar and Euclidean distance based measures perform best. We propose a factor graph model considering the sentiment-based cognitive balance theory. The results show that our model outperforms the other two well-known classifiers (logistic regression and random forest) in the mutual-follow graph and mention graphs where the strength of ties is not too weak ($@ \geq 3$). In future work, we plan to evaluate our work in a friend recommendation framework by exploiting temporal information regarding how links form.

Chapter 3

Location Estimation

In this chapter, we investigate an important application of user-location interactions—estimating the locations where messages originated by exploiting homophily principle to both social layer and geographical layer.

3.1 Introduction

With the increasing prevalent of location sharing services on social media, online content associated with a location from which it originated becomes a powerful tool in characterizing the interplay between a user's online and offline activities [Cranshaw et al., 2010]. We define a *geo-tag* as a representation of location, e.g., city, neighborhood, or latitude-longitude (lat-lon) coordinate. The geo-tagged messages provide meaningful real-time information for modeling geographical phenomena, such as monitoring regional health [Aramaki et al., 2011], detecting local emergency [Starbird et al., 2010], observing linguistic differences across geographical areas [Hong et al., 2012], and so on.

We focus on tweets in this work because of their prominence in social media and popularity over mobile devices. Although a GPS-enabled phone can geo-tag outgoing tweets, only about 2% of tweets [Leetaru et al., 2013] are. Therefore, the problem of *location estimation*, assigning a geo-tag (indicating its origin) to a tweet, is important.

Previous approaches fall into two main categories. First, content-based techniques, e.g., [Mahmud et al., 2012], assume that tweets encode location via place names or other

location words and rely on word distributions over geo-tags. They treat historical geo-tags of all users as candidates and yield large prediction errors—Section 3.4.2 revisits these. Second, individualized techniques, such as Chen et al. [Chen et al., 2013], treat a user's prior geo-tags as candidates. They map a tweet's content to its sender's interests and associate interests with locations: a user who tweets from one museum may tweet similar content from another museum. Chen et al.'s approach fails for users with insufficient historical geo-tagged tweets. Most users have sparse geo-tagged histories and some have no geo-tagged messages at all.

Our proposed technique, *Percimo* or *Personalized Community Model*, employs *geo-social communities* to overcome location data sparsity. Percimo contrasts with prior work in two ways. First, Percimo considers not only content and individual interests, but also how an individual attaches to a community, i.e., *how one's interests relate to another's locations*. Second, Percimo adopts insights from social psychology regarding attachment to a community as a basis for detecting and understanding geo-social communities.

Prentice et al.'s [Prentice et al., 1994] theory posits that a user may attach to a community in a combination of two ways: through *common bonds*—attachment to specific members of the community or through *common identity*—how much a user aligns her identity to the community, independently of its members. Sassenberg [Sassenberg, 2002] validates Prentice et al.'s theory empirically for online behavior by associating participation in "on-topic" and "off-topic" chats, respectively, with common identity and bonding. A topic serves as a seed for communal identity independent of who else is interested in that topic. When there is no fixed topic, the participants relate more to the other participants: the communal identity is weak but the bonds are strong. Grabowicz et al. [Grabowicz et al., 2013] find that communities based on interpersonal connections emphasize bonding over identity, which corroborates Sassenberg's idea.

Accordingly, we lift Sassenberg's distinction to the geo-social setting. Participation in a physical space (being near each other) indicates common identity, e.g., people living in Manhattan have a common identity. In contrast, social linkages with others indicate common bond. Percimo investigates the effect of both kinds of attachment and the synthesized the two in the location estimation, which will in turn shed light on understanding the theory. In addition, it modulates the communal aspects with personal aspects: hence

its name. Specifically, Percimo balances these aspects by assigning the most likely geo-tag to a tweet by combining historical (user's prior geo-tags, suited to a tweet about personal interests) and social (geo-tags of others in the user's community, suited to a tweet about community interests) effects.

Contributions and Main Findings. Percimo's novelty lies in how it (1) addresses data sparsity without exploding the set of candidate locations by employing communities, (2) investigates the effect of different geo-social attachment in location estimation by inspiration from sociology, specifically, the common-bond and common-identity theory, and (3) relates a user's interests to another's locations by integrating a user's personal and community interests. We evaluate Percimo via a dataset consisting of geo-tagged tweets collected over two months from two US states. We find that the synthesized attachment (bond and identity) yields least prediction error. By reducing the size of candidate sets through communities, Percimo greatly reduces the prediction error compared to a purely content-based state-of-the-art technique. By differentiating a user's community interests from personal interests, Percimo reduces prediction error over baseline models relying purely on personal history, and predicts geo-tags even for users without historical geo-tags.

3.2 Data, Problem, Framework

We evaluate our approach based on data from Twitter and Foursquare. This data includes all tweets with geo-tags in bounding boxes approximating two US states: Maryland (MD) and North Carolina (NC) from August 5 to October 8, 2013. Considering two states helps ensure geographical dispersal of users. We removed users with fewer than five tweets and tweets whose geo-tag was not lat-lon coordinates (some geo-tags are a city or neighborhood). This yielded 1,066,327 tweets from 23,897 distinct users (accounts). Using the Twitter API, we created a mutual-follow graph of users: an edge connects two users who follow each other. To mitigate sensing errors, we discretized locations into 30 m $\times$ 30 m cells on a spatial grid, generating 106,927 nonempty cells. We removed cells that were visited fewer than five times, yielding a total of 23,858 cells, each with an assigned grid ID. We posit that point of interest (POI) information provides a conceptual meaning

of a geo-tag. From Foursquare [Foursquare, 2015], we collected POIs (and each POI's top-level venue category) within a 500 m radius of each tweet's geo-tag [Chen et al., 2013]. We removed tweets with no POIs.

Let's map the general terminology to our final dataset. It contains 23,858 unique locations (grid IDs), 54,062 representative POIs, and 695,636 messages (tweets) from 12,500 users (6,824 in MD, 4,984 in NC, and 692 elsewhere). A geo-tag is a lat-lon pair; a common bond is mutually following on Twitter; and the social graph is the mutual-follow graph.

3.2.1 Problem and the Percimo Framework

Let $U = \{u\}_{u=1}^{N}$ be a set of N users and $L = \{l\}_{l=1}^{M}$ a set of M locations. Given a time T, each user has a tweet log $X_u^T = \{x_u^t\}_{t=1}^{T}$, where x_u^t represents user u's tweet at time t. A tweet x_u^t may optionally be tagged with location l_u^t representing where the tweet originated. Let $L_u^T = \{l_u^t\}$ be a set of all such locations for user u until T. And, $G(U, E)$ be a social graph, where E is the set of bonds (friendships).

Now, our research task is: Given the tweet and location log of all users until time T, social graph G, and a user u's tweet x_u^{T+1}, determine its associated location l_u^{T+1}.

$$l_u^{T+1} = \arg\max_{l \in L} P(l | X^T, L^T, G, x_u^{T+1}) \tag{3.1}$$

Figure 3.1 shows Percimo's major steps: the first two involve offline and the third step involves online processing.

Geo-social community detection involves detecting communities of users, based on both kinds of attachment (bond and identity), with similar location-visiting behavior. Communities help overcome location data sparsity by introducing geo-tags of related, as opposed to all, users.

Personal-community interest detection learns the interests relationship between users within a geo-social community from the content of their tweets.

Location estimation involves constructing a mapping $L_c^T = f(X_u^T)$ from user u's interests to location candidates, which include historical locations of u and of other users in u's geo-social community. A user's interests relate naturally to her historical locations [Chen et al., 2013; Schulz et al., 2013]. Percimo additionally relates one user's interests to *other* users' locations.

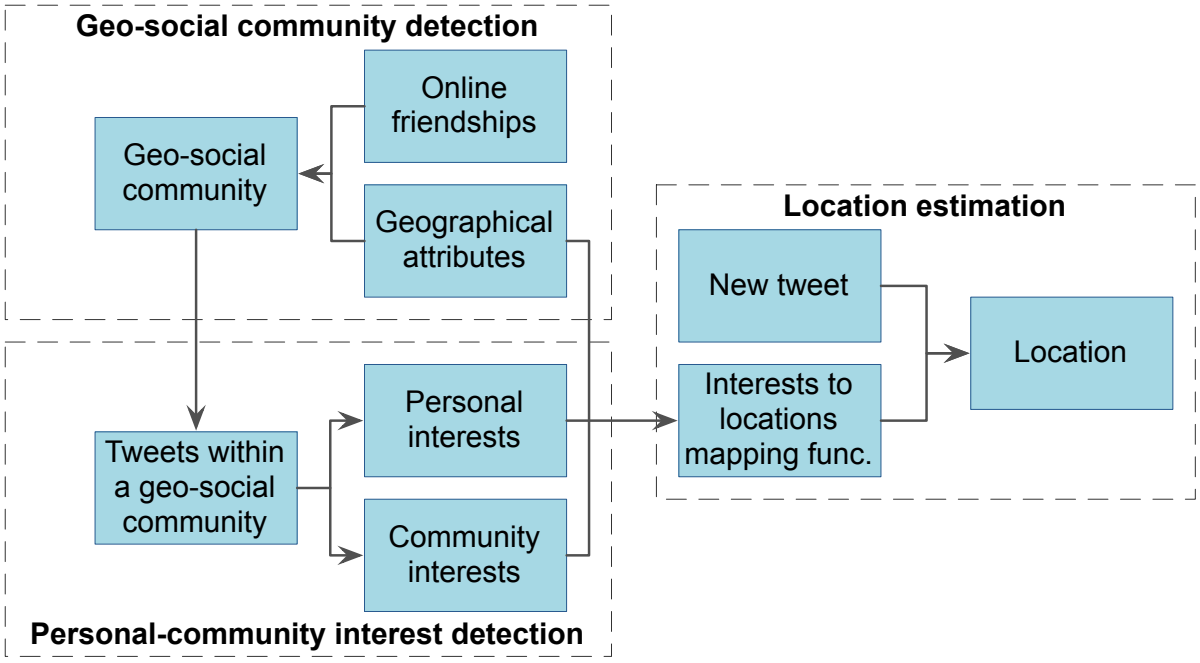

Figure 3.1 The Percimo framework.

3.3 Percimo: Proposed Approach

We now describe the three major steps of Percimo.

3.3.1 Geo-Social Community Detection

We explore three kinds of *geo-social graphs* to investigate the three corresponding geo-social attachment.

- Social (G_S): an edge between users indicates bonding.
- Local (G_L): an edge between users indicates identity (geographical distance is below a certain threshold).
- Local-social (G_{LS}): intersects social and local graphs.

We first assign to each user a representative (likeliest) location m_u. We divide users into two sets: *users with a history* (at least one prior geo-tagged tweet) and *users without*

a history (all other users). For each user with a history, the m_u is the centroid of her historical geo-coordinates L_u^T. For each user without a history, we infer her m_u via spatial label propagation [Jurgens, 2013]. Next, we compute the distance between each pair of users to decide whether the two users live locally (based on a threshold). Section 3.5 varies the threshold to investigate Percimo's prediction error.

Prentice et al. [Prentice et al., 1994] describe people forming communities spontaneously. Since acquiring ground truth on user-formed communities is not feasible, we apply a community-detection technique. We adopt Clauset-Newman-Moore [Clauset et al., 2004], but Percimo is not restricted to this algorithm.

3.3.2 Personal-Community Interest Detection

For each geo-social graph, we learn the interests of each user in the same community. We assume a tweet's content captures (some of) a user's interests. The main idea is to mimic a user's process of decision making, for example, deciding the location she wants to visit, and the set of words she wants to include in a tweet depending on her current location.

We make three assumptions about user behavior. First, a user's location visiting behavior is driven by her interests. For example, a user interested in *socializing* would go to bars whereas a user interested in *classical music* would visit a concert hall. Second, users in the same community might have similar interests (*community interests*). Third, a user's interest is based either on her *personal interest* or her community's interest. A user's interests may easily differ from her community's, especially when a community is not formed of common interests. For example, a student's interest in shopping malls may be higher than her residential hall community's, which isn't based on interest in shopping.

Our interest detection model is based on Latent Dirichlet Allocation (LDA) [Blei et al., 2003]. The input is the set of tweets of all members of a community: our model runs once for each community. We assume each tweet has only one hidden interest label—generated by either a user's personal or her community's interests, similar to a single label for each message [Chen et al., 2013; Diao et al., 2012].

Figure 3.2 shows a graphical representation of our interest detection model and Table 3.1 shows important notations. The generative process is as follows. A user u first

decides whether to go to a location from her community interests or personal interests. If she chooses the former, she selects an interest from φ_c; otherwise, she selects an interest according to η_u. With the chosen interest, words in the tweet are generated from her interest-word distribution ϕ_u. We adopt a Bernoulli distribution to indicate whether a user will choose her community interests rather than her personal interests.

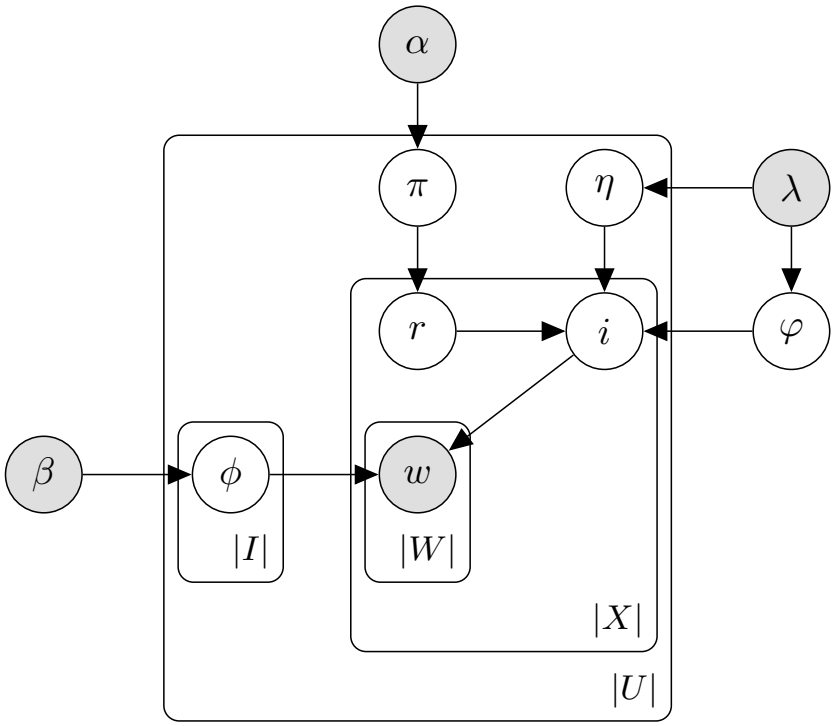

Figure 3.2 Percimo's interest-detection model.

The following steps describe the generative process:

1. For a community c, draw $\varphi_c \sim Dirichlet(\lambda)$.
2. For each user u in community c,
   a. Draw $\eta_u \sim Dirichlet(\lambda)$;
   b. For each interest, draw $\phi_u \sim Dirichlet(\beta)$;
   c. Draw $\pi_u \sim Beta(\alpha)$.

Table 3.1 Notation used in this paper

α, β, λ	Priors of Dirichlet distributions
U, L, X, R, I, W	Set of users, locations, tweets, indicators, interests, words, respectively
u, l, x, r, i, w, c	Instance of a user, location, tweet, indicator, interest, word, community, respectively
φ	Community-interest distribution
η	Personal-interest distribution
π	Bernoulli distribution over indicators
ϕ	Multinomial distribution over words
n^{-x}	The counter calculated by excluding tweet x
$n_{r,u}$	Number of times r is observed in u's tweets
$n_{i,u}$ ($n_{i,c}$)	Number of tweets by u (any user in c) that are assigned to i
$n_{w,i,u}$	Number of times that w is generated by i for u
$Y_{w,x}$	Count of word w in tweet x
Y_x	Total number of words in tweet x

3. For each tweet of a user u,
   a. Sample an indicator $r \sim Bernoulli(\pi_u)$;
   b. Sample an interest i: if $r = 1$, $i \sim Multinomial(\varphi_c)$, else $i \sim Multinomial(\eta_u)$.
4. For each word, sample $w \sim Multinomial(\phi_u)$.

The posterior probability of the latent variables in the model, given the observed data, can be factorized as follows:

$$p(\mathbf{w}, \mathbf{r}, \mathbf{i}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}) = p(\mathbf{r}|\boldsymbol{\alpha})p(\mathbf{i}|\mathbf{r}, \boldsymbol{\lambda})p(\mathbf{w}|\mathbf{i}, \boldsymbol{\beta}) \tag{3.2}$$

We adopt collapsed Gibbs sampling [Liu, 1994] to approximate the latent variables. For a tweet x, we know it is from user u. The Gibbs sampler jointly samples r_x and i_x based on the values of all other hidden variables. i_x represents u's interest for the tweet x; $\mathbf{i}^{-x}$ denotes all i except i_x; W_x denotes the set of words in tweet x (other variables

have similar symbols). For each user u, the Gibbs update equation is:

$$P(r_x, i_x|\mathbf{r}^{-x}, \mathbf{i}^{-x}, \mathbf{w}) \propto \frac{n_{r,u}^{-x} + \alpha_r}{\sum_{r \in R}(n_{r,u}^{-x} + \alpha_r)} \cdot \frac{n_{i,k}^{-x} + \lambda_i}{\sum_{i \in I}(n_{i,k}^{-x} + \lambda_i)} \cdot \frac{\prod_{w \in W_x} \prod_{y=0}^{Y_{w,x}-1}(n_{w,i,u}^{-x} + \beta_w + y)}{\prod_{y=0}^{Y_x-1}\left(\sum_{w \in W}(n_{w,i,u}^{-x} + \beta_w) + y\right)}, \quad (3.3)$$

where $k = u$ when $r_x = 0$, and $k = c$ when $r_x = 1$.

3.3.3 Location Estimation

Given a geo-social community, we now estimate the location of a new tweet by building the mapping function from users' interests to their historical locations L_c^T. Because both *historical* and *social* effects are important in modeling a user's check-in behavior in location-based social graphs [Gao et al., 2012a], Percimo seeks to integrate both effects. When a user's tweet is about her personal interests, we posit that her location is unrelated to locations of others in her community: the candidates are her historical locations (historical effect). When the user tweets about her community interests, we posit that her location may be the same as another user's: historical locations of all other users are candidates (social effect). For example, colleagues sharing interest *pizza* might go to a pizzeria for lunch on weekdays.

The probability of selecting a candidate $l \in L_c^T$ is:

$$P(l|X^T, L^T, G, x_u^{T+1}) = P(l, i_x|\eta_u, \varphi_c, \pi_u) = \mu \times P(l, i_x|\eta_u) + (1 - \mu) \times P(l, i_x|\varphi_c), \quad (3.4)$$

where $\mu \in [0, 1]$ is a parameter that controls the weight between historical and social effects. We set $\mu = P(r_x = 0|u)$, where $P(r_x = 0|u)$ is learned from π_u.

$P(l, i_x|\eta_u)$ is the probability of selecting location l from user u's historical locations. Following Chen et al. [Chen et al., 2013], we posit that a user would visit locations of a category A driven by the same interest, even if the locations are distinct. For example, for a user u, if we detected that two of her tweets are labeled with the interest *eating*, it is likely that the two tweets are sent from locations belonging to the *food* category.

$$P(l, i_x|\eta_u) = P(i_x|u)P(A|i_x, u)P(l|A, u), \quad (3.5)$$

For a user without a history, her representative location m_u (computed from label prop-

agation) is the only candidate for the historical effect, and $P(m_u, i_x|\eta_u)$ is always 1.

$P(l, i_x|\varphi_c)$ is the probability of selecting location l from user u's community's historical locations. We posit that users with the same interests and in the same community tend to visit locations with the same category, though their probabilities of visiting a location may differ.

$$P(l, i_x|\varphi_c) = P(i_x|c)P(l|i_x, c) = P(i_x|c)\frac{\sum_{v \in c} s(u, v) \times P(A|i_x, v) \times P(l|A, v)}{\sum_{v \in c} s(u, v)}, \quad (3.6)$$

where $s(u, v)$ is the similarity between users u and v. We consider only a user having history as user v.

We compute $s(u, v)$ as follows. If u has a history, we set $s(u, v)$ to be her *check-in similarity*, defined as the cosine of their check-in vectors, whose i-th component is the number of times the user visited location i [Gao et al., 2012a]. If u does not have a history, we compute $s(u, v)$ based on the distance between the representative geo-tags of the two users. Specifically, we set $s(u, v)$ to $s_{dist}(u, v) = a \times distance(m_u, m_v)^b$ [Ye et al., 2011], where $a = 0.0414$ and $b = -0.508$ are parameters set by Ye et al. [Ye et al., 2010].

3.4 Evaluation

Our objectives are to compare Percimo's prediction error (1) to that of the baseline models, (2) for three kinds of geo-social attachment, and (3) for different parameter settings.

3.4.1 Evaluation Strategy

We investigate the prediction error of Percimo on each geo-social graph. We also vary the threshold defining local users from 5 km to 40 km. Table 3.2 summarizes the statistics of the geo-social graphs we study. The subscript indicates the threshold; e.g., G_{LS_5} represents the local-social graph with the threshold 5 km. In each graph, we ignore isolated users. Since the number of users varies across graphs, to compare Percimo's prediction error across graphs, we employ the 5,623 users appearing in G_{LS_5} because these users

also appear in the other graphs. We also construct local-social graph with the threshold 5 km on the state-level sub-datasets.

Table 3.2 Statistics of the geo-social graphs

Graph	Users	Edges	Mean degree	Clustering coefficient
G_S	8,483	23,163	5.46	0.14
G_{L_5}	8,485	1,202,908	283.54	0.81
G_{LS_5}	5,623	9,350	3.33	0.19
G_{LS_10}	6,508	13,827	4.25	0.18
G_{LS_20}	7,106	16,523	4.65	0.18
G_{LS_40}	7,541	18,487	4.90	0.17
G_{LS_5} (MD)	2,930	5,441	3.71	0.17
G_{LS_5} (NC)	2,242	3,375	3.01	0.27

Parameters of Percimo. We set the total number of interests $|I|$ to 20, λ to $\frac{10}{|I|}$, and β to 0.01. We set these parameters based on guidance from previous studies [Chen et al., 2013; Hoang and Lim, 2014] and our preliminary experiments. A simple way to set α is to choose symmetric priors (i.e., $\alpha_1 = \alpha_0 = 0.5$) for each user, meaning that the user's historical locations and the locations of her community have equal influence in inferring a new location for the user. However, this may not be the case. Cho et al. [Cho et al., 2011] found that, on Brightkite (a location-based social network), there is a 53% chance that a user will check in at a location where she previously checked in, whereas only a 10% chance that she will check in at a location where a friend previously checked in. We set a user's α_1 as the user's betweenness centrality [Newman, 2010] in the subgraph of a geo-social graph induced by the user's community (and $\alpha_0 = 1 - \alpha_1$). Thus, the higher the betweenness centrality the greater the community's influence. We compute Percimo's interest-detection model for each geo-social graph via 500 iterations of Gibbs sampling. We take 25 samples with a gap of five iterations in the last 125 iterations to compute the values of all hidden variables.

We infer the representative geo-tag of a user without history via Jurgens et al.'s

[Jurgens, 2013] *geometric median select* method with 7 iterations.

Evaluation Metric. We temporally order each user's geo-tagged tweets, and take the first six weeks of data (05 August 2013 to 21 September 2013) as the training set, and test on the last two weeks of data (22 September 2013 to 08 October 2013). For each user, we predict the location of every tweet in the test set. We compare Percimo and the baseline models via average error distance (*AED*) [Chen et al., 2013]. For a tweet, error distance (*ErrDist*) is the geographical distance between the tweet's actual location and its predicted location, and the error distance of a user (*ErrDist(u)*) is the averaged error distance over all of her test tweets. Then,

$$AED = \frac{\sum_{u \in U} ErrDist(u)}{|U|} \tag{3.7}$$

3.4.2 Baseline Models

PIM (*Personal Interest Model*) [Chen et al., 2013] is most similar to Percimo among the existing works. PIM maps a user's interests detected from tweets to her historical locations and predicts the user's next location from her historical locations, not considering the social effect. We implement PIM and choose the parameters as Chen et al. do.

CM (*Content-Based Model*) Cheng et al. [Cheng et al., 2010] predict a user's location purely based on her tweets' content. We adapt this approach to consider all tweets from a given location l: $P(l|S_{words}(X_l)) = \sum_{w \in S_{words}(X_l)} P(l|w)P(w)$, where $S_{words}(X_l)$ is the set of words in all tweets from location l. We compute $P(l|w)$ via maximum likelihood estimation and $P(w)$ as $\frac{count(w)}{|W|}$, where $count(w)$ is the number of occurrences of w. We implement two enhancements Cheng et al. suggested: (1) discarding nonlocal words, and (2) performing lattice-based neighborhood smoothing.

CommPIM combines PIM and communities in geo-social graphs. We apply Chen et al.'s [Chen et al., 2013] model to detect each user's interests distribution and hidden interest label. Similar to Percimo, the location candidates are L_c^T. Whereas Percimo learns η_u from the interest detection model, CommPIM learns it from PIM: $P(l|X^T, L^T, G, x_u^{T+1}) = \mu \cdot P(l, i_x|\eta_u) + (1-\mu) \cdot \frac{\sum_{v \in c} s(u,v) P(l, i_x|\eta_v)}{\sum_{v \in c} s(u,v)}$, where $P(l, i_x|\eta_u)$ is computed according to Equation 3.5.

URLM (*User Representative Location Model*) always uses the representative location

of a user as the prediction.

CRLM (*Community Representative Location Model*) always uses the representative location of a user's community m_c as the prediction. We compute m_c by averaging the latitude and longitude of the community's users' representative locations (geo-coordinates).

3.5 Results

3.5.1 Percimo and Baseline Models

Table 3.3 shows AEDs for all models for users with and without a history, except PIM, which works only for users with a history. The type of a model indicates its main aspects: I and B for common identity and common bond, respectively and H for historical effect only (neither bond nor identity). In our dataset, 16.68% users have no history (no geo-tagged tweets in the training set). We set 5 km as the threshold defining local users (other thresholds below). On the local-social graph (G_{LS_5}), Percimo yields the least prediction error among the models compared.

Table 3.3 AEDs (km) of Percimo and baseline models

Model	Type	Users with a history	Users with no history	All users
Percimo (G_{L_5})	I	8.74	45.94	14.94
Percimo (G_S)	B	8.47	52.90	15.88
Percimo (G_{LS_5})	I+B	**6.77**	**45.02**	**13.15**
PIM	H	8.28	–	–
URLM	H	8.32	52.35	15.41
CRLM (G_{L_5})	I	12.36	46.94	18.13
CRLM (G_S)	B	63.39	79.94	66.15
CRLM (G_{LS_5})	I+B	8.90	46.37	15.15
CommPIM	I+B	7.21	46.06	13.69
CM	I	269.87	268.48	269.64

First, Percimo yields better results than PIM, suggesting that a community-based approach yields lower prediction error than individual-based approaches. Second, although Percimo and CommPIM both set μ as 1 minus a user's betweenness centrality, Percimo learns η_u via the interest-detection model. Thus, the lower prediction error of Percimo can be attributed to its interest-detection model, which effectively models the interests relationship between users, and effectively maps users' interests to their historical locations.

Among the models compared, CM's AED is worst, supporting our claim that a large candidate pool increases the probability of a tweet's predicted geo-tag to be far from the actual. Also, CM's AEDs do not differ much for the two kinds of users as CM does not consider the historical effect.

Although URLM and CRLM baselines seem naïve, their AEDs are not bad (except CRLM (G_S)), suggesting that geographical influence is a crucial factor in location estimation. Percimo and CRLM both yield their best results on G_{LS} among the three geo-social graphs. However, the common-bond attachment performs much better in Percimo. These suggest (1) the synthesized attachment performs best and (2) common-bond attachment can play an important role if we properly relate one's interests to another's locations.

3.5.2 Threshold of Defining Local Users

We vary the threshold defining local users between 5 km and 40 km to study its effect on Percimo. We restrict our analyses to the local-social graph, which has the lowest AED for both kinds of users. Figure 3.3 shows that the lower the threshold the lower the AED, in general. The AED of G_{LS_20} is higher than that of G_{LS_40} for users with a history and the reverse for users without a history. A similar pattern arises for Percimo on G_{L_5} and G_S (Table 3.3). That is, at higher thresholds, the identity effect (locality) fades, and the bonding effect (sociality) dominates.

3.5.3 Social and Historical Effects

Percimo balances social and historical effects by learning μ (Equation 3.4). Setting $\mu = 1$ and $\mu = 0$ forces Percimo to consider historical and social effect only, respectively.

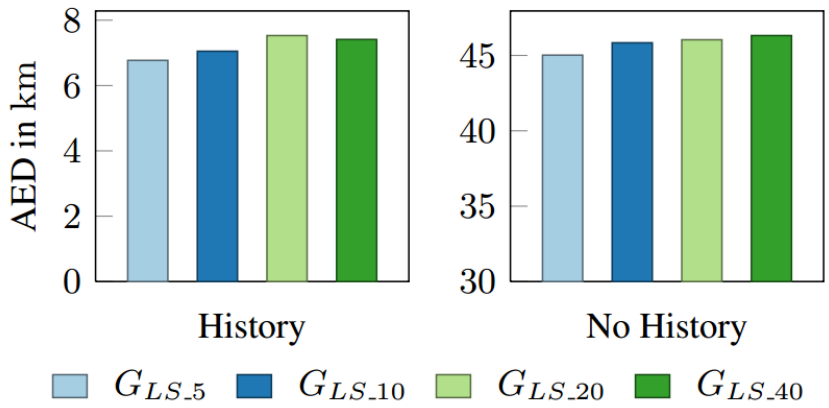

Figure 3.3 Percimo's AEDs for four local-social graphs.

Figure 3.4 compares Percimo's AEDs for the three settings of μ. The AED for $\mu = 1$ is less than that for $\mu = 0$: the historical effect is more important than the social effect for location estimation. However, Percimo's AED is least for learned μ, suggesting that both historical and social effect contribute to reducing the AED.

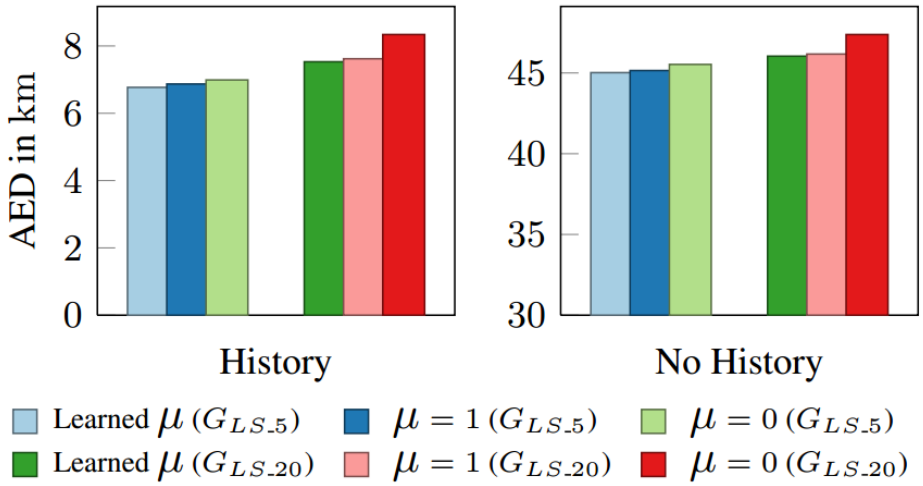

Figure 3.4 Percimo's AEDs for $\mu = 0$ (social effect), $\mu = 1$ (historical effect), and learned μ (historical & social effects).

3.5.4 Symmetric Prior vs. Betweenness Centrality

Figure 3.5 compares Percimo's AED when α_1 is set as users' betweenness centrality or 0.5. The AED is higher for $\alpha_1 = 0.5$ on both graphs, whether a user has a history or not. Thus, we conjecture that setting α_1 as users' betweenness centrality is a better choice in Percimo than setting it to 0.5 (symmetric priors). Importantly, we are not suggesting that α_1 necessarily be bound to betweenness centrality; other metrics that estimate user's attachment to her community could also be good choices. We differ this analysis to future work.

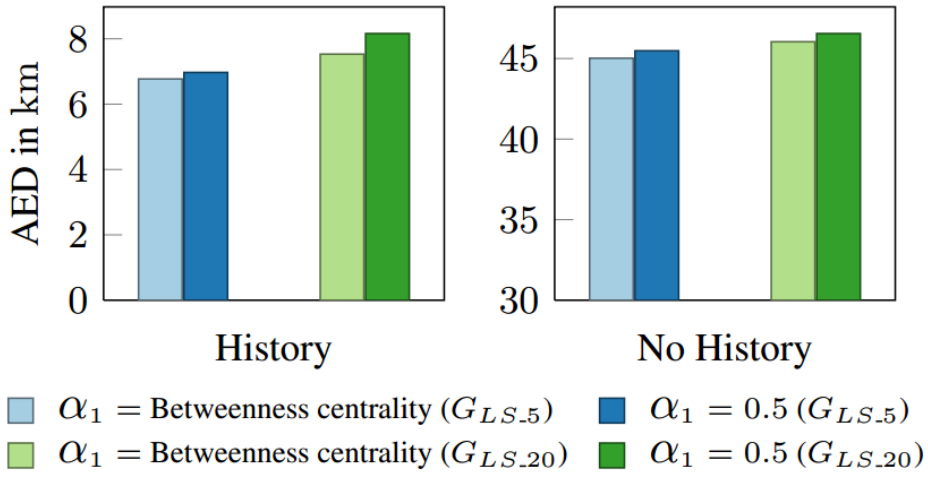

Figure 3.5 Percimo's AEDs for different α_1 settings.

3.5.5 Evaluating Percimo on State-Level Datasets

In order to test if Percimo is robust with respect to geographical scale, and investigate the relationship between performance of each model and the size of a dataset, we evaluate Percimo and other baseline models on the sub-dataset of each state. Table 3.4 presents the comparison. First, for both Maryland and North Carolina, Percimo on the local social graph G_{LS_5} performs best. Second, although the Percimo's AED is slightly lower on the

sub-dataset of Maryland, compared with that for the entire dataset, the AED of Percimo increases on the sub-dataset of North Carolina. This suggests Percimo applied in local social graph does not rely much on the size of a dataset, because the geo-social communities are detected by integrating social and geographical influence; i.e., the location candidates are always within a certain distance threshold. The same reason applies to the other models except CM.

For CM, the AED reduces greatly when we consider Maryland and North Carolina separately, suggesting that the performance of CM largely relies upon the size of a dataset. In addition, the AED on the sub-dataset of Maryland is much smaller than that on North Carolina, which might be caused by the relative sizes of the two states, Maryland being much smaller than North Carolina.

The above analyses indicate that (1) choosing an appropriate candidate pool is important in tweet localization, and (2) geographical influence can be effective in reducing prediction error and thus help make the model rely less on the size of a dataset.

Table 3.4 Comparing AEDs of Percimo and other baseline models on state-level sub-datasets

State	Model	Users with a history	Users without a history	All users
Maryland	Percimo (G_{LS_5})	**6.59**	**23.21**	**9.36**
	PIM	7.94	–	–
	URLM	6.90	31.94	11.07
	CRLM (G_{LS_5})	7.71	23.73	10.38
	CommPIM	7.43	24.07	10.20
	CM	44.38	45.61	44.58
North Carolina	Percimo (G_{LS_5})	**7.04**	**36.71**	**11.98**
	PIM	7.67	–	–
	URLM	7.23	48.37	14.09
	CRLM (G_{LS_5})	8.14	37.37	13.01
	CommPIM	7.28	39.35	12.62
	CM	111.58	120.21	113.02

3.6 Related Work

Previous work on location estimation of a user can be classified into two categories: localizing messages and localizing users. Our work falls into the first category, which estimates the location of a message based on the assumption that messages encode location-related information—either specific location names or certain words that are associated with the location where the message is sent out. Cheng et al. [Cheng et al., 2010] build a classifier for automatically identifying words in tweets with a strong geo-scope to estimate a user's city-level location. Mahmud et al. [Mahmud et al., 2012] develop an ensemble of statistical and heuristic classifiers to infer the home locations of Twitter users through their tweets and tweeting behaviors (volume of tweets per time unit). Chandra et al. [Chandra et al., 2011] extend the above approaches by employing the reply-tweet relationships between users in Twitter. Percimo differs from these works in that it estimates location at a fine-grained level; i.e., we estimate the location of every tweet.

For estimating locations at the fine-grained level, Kinsella et al. [Kinsella et al., 2011] estimate the location of every tweet by sampling the word distribution for that location. Instead of assuming the independence between words, some researchers [Priedhorsky et al., 2014; Flatow et al., 2015] model the location distributions of phrases (n-grams) and assign a location to a tweet by identifying n-grams associated with hyper-local regions. Schulz et al. [Schulz et al., 2013] propose a multi-indicator approach with dedicated location entries and user profiles. Some researchers focus on recognizing textual references to geographical locations [Lieberman et al., 2010; Li and Sun, 2014]. Instead of solving the problem by identifying spatial aspects of words in unstructured texts, Percimo addresses the problem by exploiting the correlation between users' textual content and their locations. Chen et al. [Chen et al., 2013] estimate location of a tweet by assuming that a user's interests are related to her locations. Their techniques apply to each user individually. Percimo is novel in that it exploits not only the correlation between a user's content and her locations, but also the correlation between a user's content and others' locations.

Approaches in the second category seek to predict the location of a user, not the location of a geo-tagged message. This work typically does not exploit the messages. The

location of a user could be expressed at either an abstract level (each user only has one location) or at a fine-grained level (the location of each user changes over time). Works in this category can be further classified into to two sub-categories.

First, motivated by the positive relationship between social ties and locations, several of these works claim that the locations of a user's friends are helpful in predicting the user's location. Backstrom et al. [Backstrom et al., 2010] observe that friendship between Facebook users drops monotonically as a function of distance; they use such a function to estimate the home address of Facebook users from provided addresses of their friends. Jurgens [Jurgens, 2013] develop a method of inferring users' location by spatially propagating location through social network, given a small number labeled locations. Each of these works estimates a user's location at an abstract level. Sadilek et al. [Sadilek et al., 2012] propose a probabilistic model to infer a user's find-grained location from her friends' location. They assume that each user changes her location every 20 minutes, because most users tweet with lower frequency. Gao et al. [Gao et al., 2012a] explore the contribution of social correlation in user's check-in behavior by integrating the social and historical effects. They show that both social and historical ties can help in location prediction.

The second sub-category of estimating location of a user is to mine frequent mobility patterns from the GPS trajectories. Monreale et al. [Monreale et al., 2009] build a decision tree from users' historical movement patterns, and predict the next location through finding the best matching path in the tree. González et al. [González et al., 2008] study the individual trajectory from the mobile phone data of 100,000 users. They find that human trajectories show a high degree of temporal and spatial regularity, and humans follow simple reproducible patterns. To account for the statistical characteristics of individual human trajectories, Song et al. [Song et al., 2010] build a statistically model for individual human mobility by introducing two principles: how an individual moves to a new location and how she returns to one of the previously visited locations. Cheng et al. [Cheng et al., 2011] conduct a quantitative assessment of mobility patterns from users' spatial, temporal, social, and textual aspects, and find that content analysis of posts associated with footprints can support better understanding of users' mobility movement.

In contrast, Percimo focuses on the content analysis of messages and the relationship

between a message and its associated location. Due to the large number of user-generated messages, content analysis can provide a rich source of context for estimating locations and understanding how users engage with location-based services.

Besides location estimation, Percimo is related to some techniques in POI recommendation. Many researchers exploit collaborative filtering algorithms, leveraging the similarity between users' mobility patterns and preferences for POIs, to overcome the data sparsity problem. Ye et al. [Ye et al., 2011] propose a collaborative recommendation algorithm that fuses user preference to a POI with social influence and geographical influence. Specifically, they put a emphasis on geographical influence due to the spatial clustering phenomenon exhibited in user check-in activities. Zheng et al. [Zheng et al., 2010] propose a collaborative location and activity filtering framework to find users with similar preferences and similar activity-patterns. Other researchers explore both geographical and social influence based on the recommendation framework, and they usually put a special emphasis on the special property of geographical influence. Zhang et al. [Zhang and Chow, 2013] developed a kernel density estimation approach to personalize the geographical influence on users' check-in behaviors as individual distributions rather than a universal distribution for all users. Cheng et al. [Cheng et al., 2012] modeled the geographical influence as a multicenter Gaussian model, and fused the geographical and social influence into a matrix factorization framework.

Lastly, Percimo is related to works that focusing on the community discovery and content analysis of messages. By assuming that users in a community are likely to talk about similar topics, some researchers build LDA-based models to detect such communities [Zhang et al., 2007; Henderson et al., 2010; Yin et al., 2012]. Sachan et al. [Sachan et al., 2011] propose a method of detecting communities by integrating both content of messages and link information of social graphs. Similar to Percimo, Hoang et al. [Hoang and Lim, 2014] propose a model by jointly modeling users' personal interests and community interests, but a user could participate in multiple communities. Percimo has a different underlying motivation: detecting interests of users from communities for tweet localization.

3.7 Conclusions

We estimate locations of user-generated messages such as tweets, made challenging by the sparsity of geo-tagged messages. Percimo (1) employs community structure and explores different geo-social attachment for location estimation, (2) exploits the correlation between users' textual content and locations. By reducing the candidate pool, Percimo outperforms a state-of-the-art approach that relies solely on content information. Percimo balances a user's personal and community interests to outperform a state-of-the-art technique that considers only personal interests.

Percimo's parameters affect prediction error. We find that the synthesized attachment (bond and identity) yields least AED in location estimation, and a lower threshold of defining local users could reduce the prediction error. Percimo's effectiveness is limited when a user has neither geo-tagged tweets nor social relationships though it is better than traditional approaches in this respect.

We defer modeling users' participation in overlapping and multiple communities to future work.

Chapter 4

Point-of-Interest Recommendation

In this chapter, we investigate another application of user-location interactions–POI recommendation.

4.1 Introduction

With the increasing popularity of location-based social networks (LBSNs), personalized point-of-interest (POI) recommendation has attracted attentions from both industry and academia. POI recommender systems with improved quality not only help users explore interesting places, but also benefit companies for increasing revenues.

However, the user-POI check-in matrix is highly sparse because users usually check in a few POIs. The recommendation quality is poor for approaches that rely only on the check-in matrix [Zhang and Chow, 2015]. Therefore, *context* have been exploited to improve recommendation quality [Ye et al., 2011; Cheng et al., 2012; Gao et al., 2013; Zhang et al., 2015].

Context is a multifaceted concept and its definition varies on applications. In the recommender systems, context means the location of users, the emotional status of users, purchasing purpose, and so on [Adomavicius and Tuzhilin, 2008]. In this paper, we focus on two contextual factors: *content* and *neighborhood*.

Content is serving as an important contextual factor in many applications [Liu and Xiong, 2013; Yuan et al., 2014]. We exploit content information, i.e., a user's review for a

POI and the category of a POI, to model a user's personalized preference for a POI. In the popular LBSNs, e.g., Yelp and Foursquare, users can write a review to share her opinion about the POI, which provides important and unique opportunity for us to model a user's preference toward a POI. A review typically covers several aspects of a user's comments on a POI. For example, from Alice's review about a restaurant: "Atmosphere is ok, but the pad thai here is delicious...", we can infer that she is neutral about the atmosphere, and is positive about the taste. In addition, users usually care about different aspects for POIs with different categories. For example, Alice comments service and price in her another review about a car wash store: "Great service, cheap price. My car looks great!". Existing works either ignore aspects or sentiment [Liu and Xiong, 2013; Hu and Ester, 2013; Yang et al., 2013; Gao et al., 2015]. To the best of our knowledge, only Zhang et al. [Zhang et al., 2015] jointly model aspects and sentiment on POI recommendation. However, their aspect-sentiment modeling approach is not personalized: they built a supervised framework on all users' reviews to capture the relationship between sentiment of each aspect within a review and the overall preference, ignoring the fact that different users care about different aspects of a POI. In addition, if a user doesn't have social links, her preference for an unvisited POI can't be predicted by their approach.

Neighborhood is the surrounding POIs of a POI. We claim that neighborhood is an important contextual factor in POI recommendation because of *neighborhood effect*. We define neighborhood effect as that a user's visiting behavior to a POI is not only decided by her preference to the POI, but also be affected by her preference to its nearby POIs. According to Tobler's First Law of Geography [Tobler, 1970]: "Everything is related to everything else, but near things are more related than distant things", we believe that modeling neighborhood effect is nontrivial for POI recommendation. Importantly, neighborhood effect is different from the *geographical effect* investigated by previous study [Ye et al., 2011; Cheng et al., 2012; Zhang and Chow, 2013]. Geographical effect is mainly about the cost of travel (e.g., time cost or monetary cost) from a user's current location to the POI, whereas neighborhood effect is more related to the environmental context created by the surrounding POIs. For example, a user may be more likely to visit a restaurant that is surrounded by her interested POIs than an isolated restaurant, even though the two restaurants are at the same distance from the user's current location. In

this case, the geographical effect is the same, and the user's behavior is affected by the neighborhood effect. Liu et al. [Liu et al., 2014] found that neighborhood effect is more important than geographical effect in POI recommendation. However, a lot of properties of neighborhood effect have not been investigated. For example, is it dominated by a single POI or all of the nearby POIs? Is it dominated by the least or most preferred POI?

Overall, we are interested in three research questions:

- How can we model a user's personalized preference for a POI via her aspect-based sentiment?
- How can we model neighborhood effect on POI recommendation?
- How can we exploit both content-based preferences and the neighborhood effect to improve the quality of POI recommendation?

To answer the first question, we propose an unsupervised method to learn a user's category-based aspect distribution. The aspect distribution of a POI is scaled by the aspect-based sentiment of users who have visited it. We profile both users and POIs via the aspect distributions, and model a user's content-based preference toward a POI by calculating the similarity between the two distributions. For the second question, we discretize POIs into cells on a spatial grid with a certain threshold, and propose different sets of features that capture different properties of a neighborhood. Finally, we fuse geographical influence with the content-based preference modeling and neighborhood effect to develop a POI recommendation framework. Our evaluation through a Yelp dataset demonstrates that (1) exploiting aspect-based sentiments could effectively improve the performance of preference modeling, (2) neighborhood effect does exist when a user decides her POI visiting behavior; 500 m is a better choice than 200 m to define a neighborhood to exploit neighborhood effect; the neighborhood properties features (e.g., average visit) are more effective in modeling neighborhood effect than features that are based on a user's preference toward nearby POIs individually.

4.2 Data and Problem Definition

We evaluate our approach on Yelp Challenge Dataset [Yelp, 2015a]. Even though the dataset covers several cities, majority of them contain less than 1,000 POIs. To overcome the data sparsity problem, we only consider POIs in Phoenix, which is the city that contains most POIs. We remove both users and POIs with fewer than five reviews, this yields 129,020 reviews written by 8,557 unique users for 5,818 unique POIs. Each POI is associated with a latitude-longitude coordinate (geo-tag). and a top-level category [Yelp, 2015b]. The resulting dataset covers POIs from 22 different categories.

4.2.1 Problem Definition

Let's first formalize definition of terms in this paper:

Definition 1 (Aspect) *An aspect is an attribute of a POI.*

For example, the aspects of a restaurant could be "price", "service", "atmosphere", and so on.

Definition 2 (Sentiment) *In our paper, a sentiment contains three real-valued scores, indicating its positivity, negativity, and objectivity, and the sum of the three scores is one. A sentiment is associated with an aspect. We say a user's sentiment is positive if her positivity score is larger than negativity score, and vice versa.*

Let $U = \{u\}_{u=1}^{M}$ be a set of M users and $L = \{l\}_{l=1}^{N}$ a set of N POIs. We use locations and POIs interchangeably. Each POI is associated with a geo-tag and a category. We also have the user-POI check-in information, representing the number of times that a user has visited a POI. Additionally, each user has written a review for the POIs that she has visited. Our paper address the following three problems:

Problem 1 (Preference Modeling via Aspect-Based Sentiment) *Given the observed reviews from all users and the category of each POI, the goal is to (1) extract a user's sentiment toward different aspects for different POIs, (2) profile each user and each POI in terms of a aspect distribution, (3) obtain an aspect-matching score to estimate how well a POI meets a user's expectation.*

Problem 2 (Neighborhood Effect Modeling) *Given the aspect-matching scores between users and POIs, users' observed check-in information, the goal is to model neighborhood effect when a user decides her POI visiting behavior, by investigating different properties of the neighborhood effect.*

Problem 3 (POI Recommendation) *With the joint efforts of preference modeling and the neighborhood effect, the goal is to predict the preference score from a user to an unvisited POI, and return the top-k POIs with the highest scores.*

4.3 Preference Modeling via Aspect-Based Sentiment

In this Section, we describe a *content filtering* based approach to understand users' preferences for POIs. We profile a user via a category-based aspect distribution in order to capture her interested aspects. We assume that a user usually cares about different aspects when she visits POIs belonging to different categories. For example, a user typically cares about *taste* and *service* than other aspects when she decides to visit a POI belonging to *food* category. Similarly, we profile a POI via an aspect distribution in order to capture its nature from users' comments, e.g., *taste* is good but *price* is high. These profiles allow us to model how well a POI would match a user's expectation. In the context of recommendation problem, where a user has no historical interaction with an unvisited POI, exploiting content information can address the data sparsity problem to some extent. In addition, the profiles built from content information not only capture many significant qualities of a user or a POI, but also offer an interpretable way to understand users' relevance to POIs.

Our preference modeling module is composed of three steps: extracting sentiment, profiling users and POIs in terms of aspect distributions, and obtaining user-POI aspect matching scores.

4.3.1 Sentiment Extraction

We split each review into sentences, and each sentence is considered as an aspect.

First, we use Stanford named entity recognizer [Finkel et al., 2005] to replace named entities with the corresponding symbols (e.g., replace "$21.49" with #MONEY# and

"New York" with #location#). In addition, we replace a URL with #LINK#. Second, we extract nouns and adjectives from each sentence using Stanford Part-Of-Speech Tagger [Toutanova et al., 2003]. Next, we use Porter's [Porter, 1980] stemmer algorithm on each noun and adjective. Lastly, we handle negation in a way that if negation is found before nouns or adjectives, we add "non_" before the word. After processing, the words in each sentence are divided into two parts: *aspect words* (nouns) and *sentiment words* (adjectives). Following Zhang et al.'s approach [Zhang et al., 2015], we only consider nouns representing aspects. Adjectives are strong indicators of sentiment [Hatzivassiloglou and Wiebe, 2000], and considering adjectives alone can improve sentiment prediction accuracy [Bakliwal et al., 2013].

Next, we use the sentiment words of each sentence to obtain each user's sentiment through an established sentiment lexicon: SentiWordNet [Baccianella et al., 2010]. In SentiWordNet, each word is associated with three real-valued scores, indicating its positivity, negativity, and objectivity, and the sum of the three scores is one. We choose the adjectives in the lexicon and stem them to build a pairwise stem-score mapping dictionary. For each aspect, we compute its sentiment in the following way: first, we obtain the sentiment of each sentiment word from the dictionary; if the sentiment word starts with "non_", we exchange its positivity and negativity scores. We then average positivity, negativity, and objectivity scores for each sentiment word to obtain the sentiment of an aspect.

4.3.2 Users Profiling

To profile a user, we group all the reviews that she has written and detect her category-based aspect distribution. The modeling process has two assumptions. First, for POIs belonging to different categories, a user is interested in different aspects. For example, a user may care more about *taste* and *price* for a restaurant whereas care more about *service* and *location* for a hotel. Second, the more a user cares about an aspect, the more comments she makes about the aspect, no matter whether her sentiment toward the aspect is positive or not. For example, if a user likes the taste in a restaurant whereas dislikes the taste in another, we can infer that she is interested in *taste* for POIs belonging to the *food* category, even though her sentiments are opposite.

We develop our aspect detection model based on Latent Dirichlet Allocation (LDA) [Blei et al., 2003]. The input is the *aspect words* from all the reviews. Each sentence of a review has one hidden aspect label. Figure 4.1 is the graphical representation of our aspect detection model, and Table 4.1 summarizes important notations. The generative process is straightforward: if a user u visits a POI belonging to the category c, her aspects in the review are generated from her category-based aspect distribution $\varphi_{u,c}$. Specifically, for each sentence, she first chooses an aspect z. With the chosen aspect, words in this sentence are generated from the aspect-word distribution ϕ_z.

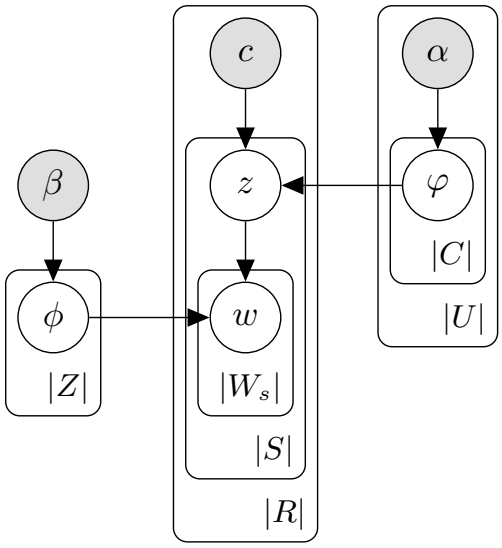

Figure 4.1 A graphical representation of our aspect detection model.

1. For each aspect z, draw $\phi_z \sim Dirichlet(\beta)$.
2. For each user u,
   (a) For each category c, draw $\varphi_{u,c} \sim Dirichlet(\alpha)$.
3. For each review with a category c from a user u,
   (a) For each sentence, sample an aspect $z \sim Multinomial(\varphi_{u,c})$,
      i. Sample each word $w \sim Multinomial(\phi_z)$.

Table 4.1 Notation used in the aspect detection model

α, β	Priors of Dirichlet distributions
U, L, R, S Z, W, C	Set of users, POIs, reviews, sentences, aspects, words, categories, respectively
u, l, r, s z, w, c	Instance of a user, POI, review, sentence, aspect, word, category, respectively
φ ϕ	Personal aspect distribution Multinomial distribution over words
n^{-s} $n_{z,c,u}$ $n_{w,z}$	The counter calculated by excluding s Number of times z is observed in u's reviews about POIs belonging to c Number of times that w is generated by z
$Y_{w,s}$ Y_s	Count of word w in sentence s Total number of words in sentence s

We factorize the joint probability of $\mathbf{z}$ and $\mathbf{w}$ as follows:

$$P(\mathbf{w}, \mathbf{z}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{c}) = P(\mathbf{z}|\boldsymbol{\alpha}, \mathbf{c})P(\mathbf{w}|\mathbf{z}, \boldsymbol{\beta}) \tag{4.1}$$

We use collapsed Gibbs sampling [Liu, 1994] to sample hidden variables and estimate the model parameters from these samples. For a sentence s in a review, we know it is from user u and it is about category c. The Gibbs sampler samples the aspect assignment z_s based on the aspect assignments of all other sentences, where z_s represents the aspect of sentence s. Let $\mathbf{z}$ denote the set of all hidden variables z and $\mathbf{z}^{-s}$ denote all z except z_s. In addition, W_s denotes the set of words in sentence s. We use similar symbols for other variables. The Gibbs sampling equation is:

$$P(z_s|\mathbf{z}^{-s}, \mathbf{w}) \propto \frac{n^{-s}_{z,c,u} + \alpha_z}{\sum_{z \in Z}\left(n^{-s}_{z,c,u} + \alpha_z\right)} \cdot \frac{\prod_{w \in W_s}\prod_{y=0}^{Y_{w,s}-1}\left(n^{-s}_{w,z} + \beta_w + y\right)}{\prod_{y=0}^{Y_s-1}\left(\sum_{w \in W}\left(n^{-s}_{w,z} + \beta_w\right) + y\right)}, \tag{4.2}$$

The learned category-based distribution $\varphi_{u,c}$ is the profile of each user. If a user hasn't visited any POI for a certain category, her aspect distribution of the category is the prior

distribution with the chosen α.

4.3.3 POIs Profiling

To profile a POI, we group all users' reviews about the POI to obtain its aspect distribution. However, the modeling process is a little bit different from that of users. Specifically, except the aspect distribution of each review, we also need users' sentiment toward these aspects. For example, if most users comment positively about a restaurant's *taste*, whereas comment negatively about its *service*, the restaurant won't match a user's expectation well if the user cares more about *service* than *taste* when she visits a POI belonging to the *food* category. Therefore, we should use a user's sentiment to scale the importance of each aspect for profiling POIs.

Given a sentence s written by user u for POI l, its scaling factor is computed according to the following Equation:

$$\eta_{s,l,u} = 1 + p_{s,l,u} - n_{s,l,u} \tag{4.3}$$

where $p_{s,l,u}$ and $n_{s,l,u}$ are obtained in Section 4.3.1, representing the sentence's positivity and negativity score respectively.

With the estimated parameters in Section 4.3.2, we derive each POI's aggregate aspect distribution π_l with sentiment scaling factors. For a POI l, the probability of aspect z is computed as follows:

$$\frac{(\sum_{u \in U} i_{z,s,l,u} * \eta_{s,l,u}) + \alpha_z}{\sum_{z \in Z} \left((\sum_{u \in U} i_{z,s,l,u} * \eta_{s,l,u}) + \alpha_z \right)} \tag{4.4}$$

where $i_{z,s,l,u}$ is an indicator: it equals to 1 if sentence s written by user u for POI l is assigned to aspect z, otherwise $i_{z,s,l,u} = 0$.

4.3.4 Aspect Matching Score

For each POI with category c, it is profiled via an aspect distribution, and each user has an aspect distribution for a given category. With these two distributions, we can obtain an matching score that models how well a POI matches a user's expectation. For example, if a user cares about both *service* and *taste* for POIs belonging to *food* category; a restaurant has been positively commented for both aspects, whereas another

restaurant's reputation of *service* is bad, the first restaurant should match better given the user's expectation, and its aspect matching score is higher accordingly.

We use the symmetric Jensen-Shannon divergence to compute the distance between user u's category-based aspect distribution $\varphi_{u,c}$ and POI l's aspect distribution π_l:

$$D_{JS}(u,l) = \frac{1}{2}D(\varphi_{u,c}\|M_{u,l}) + \frac{1}{2}D(\pi_l\|M_{u,l}) \tag{4.5}$$

where $M_{u,l} = \frac{1}{2}(\varphi_{u,c} + \pi_l)$, and $D(\cdot\|\cdot)$ is the Kullback-Leibler divergence [Kullback and Leibler, 1951]. Therefore, the aspect matching score between a user u and a POI l is:

$$a_{u,l} = 1 - D_{JS}(u,l) \tag{4.6}$$

4.4 Neighborhood Effect Modeling

To motivate the neighborhood effect, we calculate the geographical distance between every two POIs by the Haversine formula [Sinnott, 1984]. In our dataset, 80%, 91%, 97% POIs have at least one POI next to it within 100 m, 200 m, 500 m, respectively. Therefore, most POIs are not geographically independent. In addition, according to Tobler's First Law of Geography [Tobler, 1970]: "Everything is related to everything else, but near things are more related than distant things", we claim that a user's preference toward a POI is not independent of her preference toward its nearby POIs.

Our next goal is to model neighborhood effect when a user decides her next visiting POI. Specifically, except aggregating a user's preference over each nearby POI [Zhang et al., 2015], we are interested in that whether there are other ways of modeling neighborhood effect, and what's the performance of each way on POI recommendation.

The following are the notations in this Section:

- Let $n_{u,l}$, $n_{u,c}$ be the number of times that user u visits POI l, POIs belonging to category c, respectively.
- Let $a_{u,l}$ be the aspect matching score between user u and POI l.
- Let $r_{u,l}$ be the sum of rating that user u gives to POI l (a user may rate a POI several times).
- Let U_l, U_c be the set of users who have visited POI l, POIs belonging to category c, respectively.

- Let L_n be the set of POIs, except POI l, that locate in the same neighborhood with l.
- Let $L_{n,c}$ be the set of POIs belonging to category c, except POI l, that locate in the same neighborhood with l.
- Let U_n be the set of users who have visited at least one POI (not including POI l) in the neighborhood where l locates.

4.4.1 User-POI Preference

Before modeling the neighborhood effect, we define features modeling a user's preference toward a single POI.

Because the key of POI recommendation is to estimate a preference score between a user and her unvisited POI, a lot of features would be unknown, e.g., rating, number of visits. However, we can still obtain a user's preference for a category of a POI and how well a POI will match her expectation in terms of the aspect matching score, even though there is no direction interaction between the user and the POI.

To model a user's preference for a category, two intuitions should be considered. First, the category captures a user's interests when she decides to visit a POI. For example, a user will go to bars if she is interested in *socializing*; or she will go to gyms if she is interested in *doing exercises*. Second, if a user visits a category that is rarely visited by others, the user would more be interested in this category. For example, in the Yelp dataset, POIs belonging to the *food* category are visited more often than others, which does not imply that everyone is most interested in *eating*. If a user visits *concert halls* more frequently than other users, it is likely that she is very interested in *classical music*.

The term frequency-inverse document frequency (TF-IDF) [Salton and Buckley, 1988] is a metric that captures both of the intuitions. In our context, category is the term and a user's check-in history is a document. Therefore, we define a user's preference for a category c as:

$$b_{u,c} = \frac{n_{u,c}}{\sum_{c \in C} n_{u,c}} \times \log \frac{|U|}{|U_c|} \tag{4.7}$$

Zhang et al. [Zhang and Chow, 2015] measures a user's preference toward a POI in terms of the POI's popularity weighted by the user's category preference. Enlightened by their definition, we first define a POI's popularity by the following metrics:

- Number of visits: $n_l = \sum_{u \in U_l} n_{u,l}$
- Average rating: $r_l = (\sum_{u \in U_l} r_{u,l})/n_l$
- Number of users who have visited the POI: $|U_l|$

Overall, given a user u and a POI l with category c, her preference toward the single POI can be modeled as:

User-POI Features: we introduce a parameter τ to overcome the "cold start" problem [Herlocker et al., 2004], where a user does not have sufficient history for a certain category.

- Aspect matching score (ASPECT-MATCHING): $a_{u,l}$
- TF-IDF preference toward category (CATEGORY-PREFERENCE): $b_{u,c}$
- Categorical popularity by number of visits (VISIT-POPULARITY): $(b_{u,c} + \tau) \times \frac{n_l}{\sum_{l \in C} n_l}$
- Categorical popularity by rating (RATING-POPULARITY): $(b_{u,c} + \tau) \times \frac{r_l}{\sum_{l \in C} r_l}$
- Categorical popularity by number of users (USER-POPULARITY): $(b_{u,c} + \tau) \times \frac{|U_l|}{\sum_{l \in C} |U_l|}$

The user-POI features are not limited to the above five features.

4.4.2 User-Neighborhood Preference

To define a neighborhood, we discretize POIs into cells on a spatial grid according to a certain threshold, and POIs within a cell belong to a neighborhood. We choose three different thresholds: 100 m × 100 m, 200 m × 200 m, 500 m × 500 m. If a person's walking speed is 5 km/h [Wikipedia, 2016], 500 m is a 6-minute walk distance.

We define the size of a neighborhood as the number of POIs in the neighborhood. Figure 4.2 outlines number of neighborhoods in terms of different neighborhood sizes with the three thresholds. We can observe that all of the three distributions fit a power-law distribution: majority of the neighborhoods have less than 20 POIs, and only a few neighborhoods consist of more than 20 POIs.

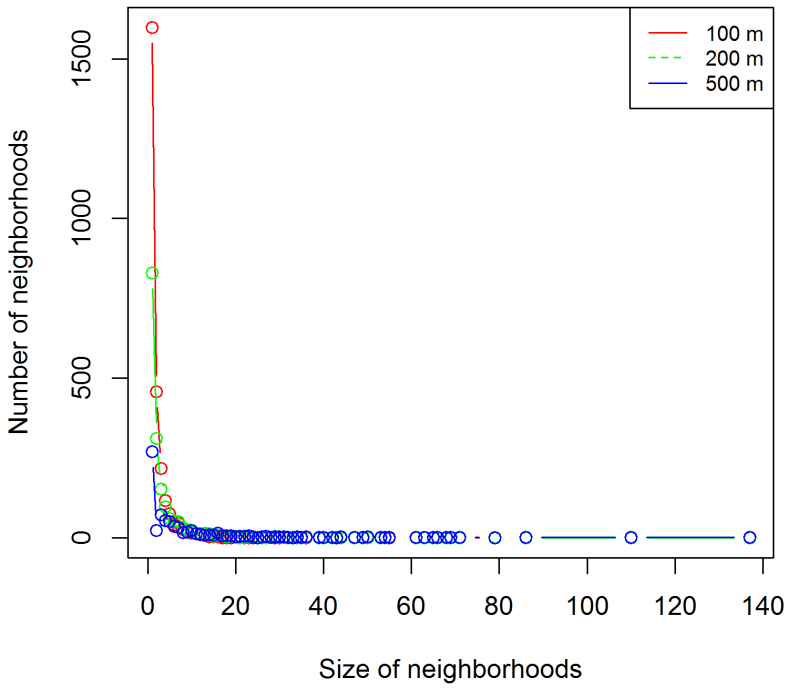

Figure 4.2 Number of neighborhoods in terms of different neighborhood sizes, within threshold of 100 m, 200 m and 500 m.

For a user u and a POI l, we define u's preference for the neighborhood where l locates by ignoring u's preference for l; we want to investigate how the l's nearby POIs affects u's preference for l, hence we call it *neighborhood effect*. Intuitively, the neighborhood effect comprises two factors: the user's preference toward the each nearby POI individually (personalized features) and the neighborhood property (property features). Accordingly, we propose two set of features:

Personalized Features: to investigate different properties of the neighborhood effect, we divide personalized features into four subsets. The first two subsets aggregate a user's personalized preference over each nearby POI, whereas the latter two subsets model the neighborhood effect via a single nearby POI (most or least preferred nearby

POI from a user).

To simplify the notation, let $f_{u,j}$ represents one of the user-POI features. Thus each one of the following subsets is replaced with five features in evaluation. In addition, $d_{l,j}$ is the geographical distance between POI l and POI j.

- Average preference over each nearby POI
  (AVERAGE-SCORE): $(\sum_{j \in L_n} f_{u,j})/|L_n|$,
- Average preference over each nearby POI, but the weight is inversely proportional to the distance
  (DISTANCE-SCORE): $(\sum_{j \in L_n} f_{u,j} \times \frac{1}{d_{l,j}})/(\sum_{j \in L_n} \frac{1}{d_{l,j}})$,
- Maximum preference among nearby POIs
  (MAX-SCORE): $\max_{j \in L_n} f_{u,j}$
- Minimum preference among nearby POIs
  (MIN-SCORE): $\min_{j \in L_n} f_{u,j}$

Property Features: features belonging to this sets capture the environmental context created by nearby POIs. For example, is the neighborhood a *food plaza* or a *shopping mall*? Is it popular?

- CATEGORY-ENTROPY: $-\sum_{c \in C} \frac{|L_{n,c}|}{|L_n|} \times \log \frac{|L_{n,c}|}{|L_n|}$
- NUMBER-OF-POI: $|L_n|$
- AVERAGE-VISIT: $(\sum_{j \in L_n} n_j)/|L_n|$
- AVERAGE-RATING: $(\sum_{j \in L_n} r_j)/|L_n|$
- NUMBER-OF-USER: $|U_n|$

4.5 Evaluation of Modeling Content-based Preference and Neighborhood Effect

Before we propose our recommendation model, we evaluate the two contextual factors. Specifically, we seek to answer two questions:

- Does our content-based preference modeling approach benefit POI recommendation?

- What's the performance of different ways of modeling the neighborhood effect on POI recommendation?

Evaluation Strategy. Because the geographical effect is very important in the POI recommendation, we remove the geographical effect in this Section to avoid its interference in evaluating the content-based preference modeling and the neighborhood effect. First, we assign a representative (likeliest) location to each user. We compute the representative location by averaging the latitude-longitude coordinates of her visited POIs. However, it can also be obtained from other sources such as user profiles in real application. Second, for each user, we compute the distance between her representative location and each of her visited POIs, and average the distances to estimate her activity distance. Figure 4.3 outlines the cumulative number of users whose activity distances are less than a certain distance. We found that 90% $(7,461)$ users' activity distances are less than 10 km. In addition, the driving time of 10 km is around eight minutes if the driving speed is 72 km/h. Therefore, we assume that within 10 km, the geographical distance is not a concern for users' POI visiting behavior.

Next, for each user, we apply five-fold cross validation. That is, we randomly split each user's visited POIs into five subsets. In each round, 20% of a user's visited POIs serve as the test set, and her rest visited POIs is treated as the training data. We report the results of every round as well as the average results. In addition, for each user, we also sample a set of unvisited POIs whose number is the same with her visited POIs, and 80% of them serve as the training data. The unvisited POIs of a user are within 10 km from her representative location. In this way, we transform the recommendation problem to a regression problem. A pair is constructed between a user and a POI, and the label of the pair is the number of times that the user has visited the POI. If a user hasn't visited a POI, the label is zero. The features of each pair are computed based on the knowledge from the training data. We apply linear regression and M5 decision trees (M5 Tree) [Quinlan, 1992] from the WEKA framework with the default parameters. All the features are normalized to $[0, 1]$ before evaluation.

Parameters Setting. We set the total number of aspects $|Z|$ to 30, α to $\frac{50}{|Z|}$, and β to 0.01. We set these parameters based on guidance from previous studies [Diao et al., 2012] and our preliminary experiments. We run our model via 500 iterations of Gibbs

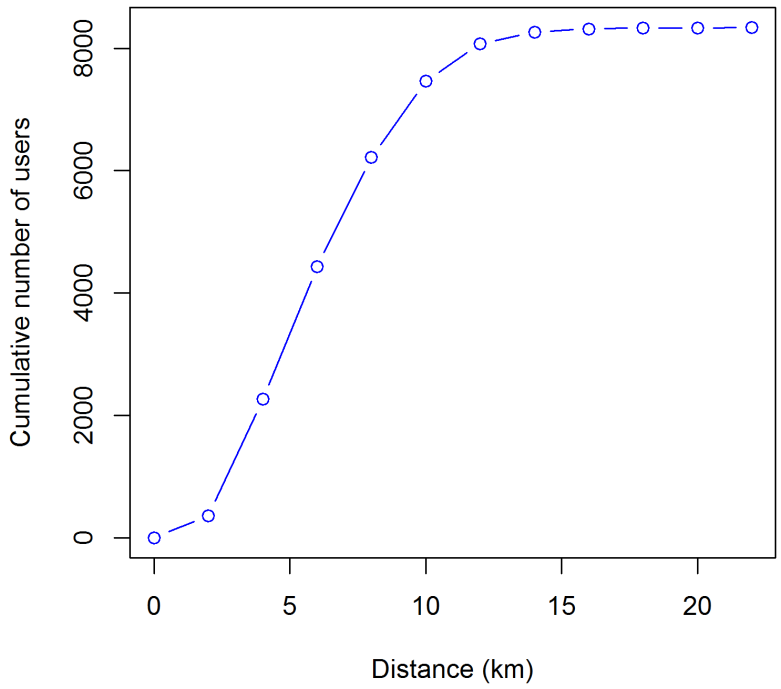

Figure 4.3 Cumulative number of users whose activity distances are less than a certain distance.

sampling. We take 40 samples with a gap of five iterations in the last 200 iterations to compute the values of all hidden variables.

Evaluation Metric. We adopt the Root Mean Squared Error (RMSE) as a evaluation metric. In our scenario, RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{K}\sum_{u,l}(\hat{n_{u,l}} - n_{u,l})^2} \tag{4.8}$$

where $\hat{n_{u,l}}$ is the predicted number of visits from a user to a POI, and K is the number of samples in the test set. A smaller RMSE indicates a more precise prediction.

For simplicity, we write the *Personalized Features* and *Property Features* as *neighborhood features*, and the *user-POI Features* as *POI features*.

4.5.1 Results of Content-based Preference Modeling

Table 4.2 shows representative aspects with top ranked words discovered by our aspect detection model. We manually assign a name to these aspects.

Table 4.2 Representative aspects with top ranked words discovered by our aspect detection model.

Aspect	Ranked words
Service	servic custom staff owner time peopl employe manag place employe
Entrée	salad chicken appet side steak dish sandwich shrimp potato plate
Price	#money# price meal deal menu special tip bill portion food
Drink	beer wine drink glass bottl cocktail bar food margarita price
Dissert	cream chocol dessert ic cake flavor cooki butter cupcak bread
Pizza	pizza chees sauc crust bread salad pasta top flavor slice
Parking	park lot #location# locat place street airport valet car spot
Atmosphere	patio tabl room seat bar area place atmosphe wall decor
Salon	hair time nail salon #person# massag pedicur cut appoint color

Next, we evaluate the effect of feature ASPECT-MATCHING; we want to evaluate whether it can improve the performance of preference modeling from a user to a POI (POI features). Figure 4.4 compare the RMSE scores with ASPECT-MATCHING (all POI features) and without ASPECT-MATCHING. With both models, RMSE scores decrease after adding ASPECT-MATCHING for each fold, even though the reduction is less with M5 Tree model. To investigate whether ASPECT-MATCHING indeed help reduce the RMSE score, we conduct a paired t-test, and each pair is the results with the same fold computed by the same model (linear regression or M5 Tree): each value in sample one is the RMSE score with ASPECT-MATCHING; each value in sample two is the RMSE score without ASPECT-MATCHING. The p-value is 0.0042, indicating the difference is statistically significant. Therefore, we can conclude that our preference modeling approach of exploiting aspect-based sentiment could effectively improve the performance of POI recommendation.

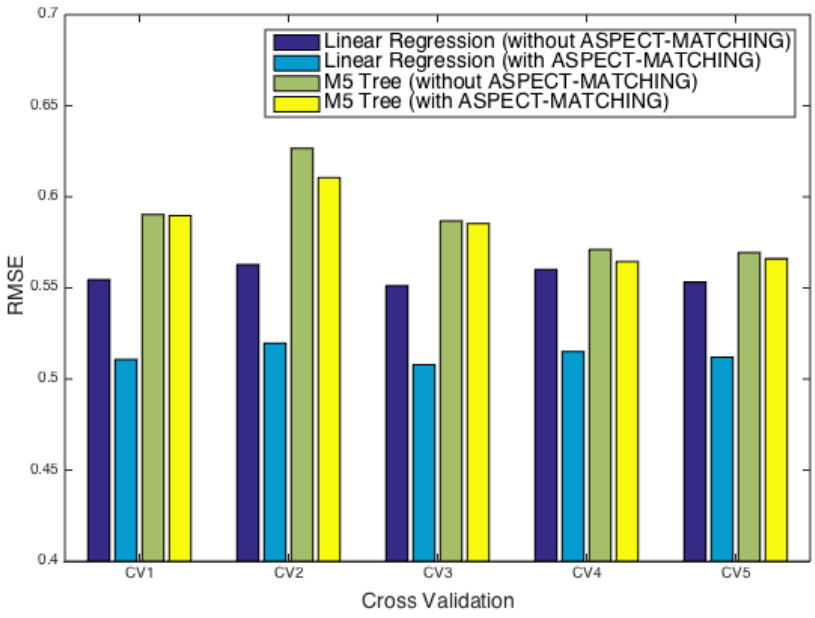

Figure 4.4 RMSE scores with and without ASPECT-MATCHING for user-POI preference modeling.

4.5.2 Results of Neighborhood Effect Modeling

Table 4.3 compares the RMSE scores with and without neighborhood features. Because POI features play a major role in POI recommendation, all the evaluations are done with the POI features. By adding neighborhood features, with the 100 m threshold, the average RMSE increases with the linear regression model but decreases with M5 Tree model; if the threshold increases to 200 m or 500 m, the average RMSE reduces with both models. Even though the average RMSE is much less in the 200 m-neighborhood than that in the 500 m-neighborhood with M5 Tree, RMSE reduces in each of the five rounds in the 500 m-neighborhood with both models, compared with POI features only. To investigate whether RMSE scores could be reduced significantly by adding neighborhood features, we conduct three paired t-test. Each test studies the effect of a certain neighborhood threshold: each value in sample one is the RMSE score with POI features only; each value in sample two is the RMSE score with neighborhood features with the threshold.

The p-value is 0.9609, 0.2881, 0.0004, corresponds with the threshold 100 m, 200 m, 500 m, respectively. Therefore, we claim that the neighborhood effect exists when a user decides her POI visiting behavior. However, the threshold of defining a neighborhood matters; in general, the threshold should not be too small (e.g., 100 m). We claim that 500 m is a reasonable choice for defining neighborhood threshold because (1) RMSE scores significantly decrease, (2) a threshold that is larger than 500 m exceeds the walking distance of a person, and the neighborhood effect would fade or even disappear.

Table 4.3 RMSE scores with and without neighborhood features.

Model	Cross Validation	POI	N (100 m)	N (200 m)	N (500 m)
Logistic Regression	cv1	0.5107	0.5092	0.5095	0.5003
	cv2	0.5197	0.5368	0.5199	0.5076
	cv3	0.5078	0.5063	0.5000	0.5070
	cv4	0.5151	0.5278	0.5049	0.5031
	cv5	0.5119	0.5360	0.5124	0.5095
	Average	0.5130	0.5232	0.5093	0.5055
M5 Tree	cv1	0.5896	0.5738	0.5843	0.5793
	cv2	0.6106	0.4192	0.6053	0.5955
	cv3	0.5854	0.4204	0.4201	0.5803
	cv4	0.5644	0.7127	0.5694	0.5562
	cv5	0.5660	0.7235	0.5704	0.5613
	Average	0.5832	0.5699	0.5499	0.5745

Neighborhood Features. To investigate how to properly model neighborhood effect, we further evaluate each set of neighborhood features. Because the RMSE is smaller with linear regression model in Table 4.3, we only report results with linear regression model in the 500 m neighborhood. Table 4.4 presents the RMSE on each set of neighborhood features for each round of cross-validation. Again, all the evaluations are done with the POI features. We can observe that (1) the property features significantly reduce RMSE more than personalized features. The p-value of a paired t-test for RMSE scores between property and personalized features is 0.0476. (2) there is not much difference

among different subset of personalized features in terms of RMSE scores. We conduct a one-way ANOVA test over the subsets of personalized features: RMSE scores with one subset of personalized features belong to a group. Thus, there are four groups for the ANOVA test, and each group comprises five RMSE scores. The p-value is 0.987.

Table 4.4 RMSE score on each set of neighborhood features with Linear Regression Model.

Cross Validation	Property	Personalized	Average	Distance	Max	Min
cv1	0.5094	0.5118	0.5110	0.5113	0.5111	0.5106
cv2	0.5083	0.5191	0.5196	0.5196	0.5139	0.5196
cv3	0.5065	0.5080	0.5082	0.5079	0.5082	0.5078
cv4	0.5042	0.5148	0.5147	0.5147	0.5158	0.5151
cv5	0.5106	0.5113	0.5113	0.5113	0.5120	0.5119
Average	0.5078	0.5130	0.5130	0.5130	0.5122	0.5130

4.6 POI Recommendation Model

4.6.1 Recommendation by Matrix Factorization

Matrix factorization is the most popular and widely used technique for recommendation [Koren et al., 2009]. The basic idea of matrix factorization is to factorize both users and POIs into a shared space with dimension $k \ll min(M, N)$, and $U_i \in \mathbb{R}^{1\times k}$ and $L_j \in \mathbb{R}^{1\times k}$ represent latent factors of a user and a POI, respectively. Let U and L denote the set of latent factors of all users and all POIs, respectively; U_i is the i^{th} row in U and L_j is the j^{th} row in L. The preference of a user for a POI can be predicted as

$$\hat{R_{i,j}} = U_i L_j^\intercal \tag{4.9}$$

Overall, the basic POI recommendation model learns U and L by solving the following optimization problem:

$$\min_{U,L} \frac{1}{2} \sum_{i,j} I_{i,j}(R_{i,j} - \hat{R_{i,j}})^2 + \frac{\lambda_1}{2}\|U\|_F^2 + \frac{\lambda_2}{2}\|L\|_F^2 \tag{4.10}$$

where $I_{i,j} = 1$ if user i has checked in at POI j, otherwise $I_{i,j} = 0$. $\|\cdot\|_F^2$ is the Frobenius norm of a matrix; λ_1 and λ_2 are regularization parameters. Once U and L have been learned, the preferences between a user and any of her unvisited POIs can be obtained by Equation 4.9

4.6.2 Incorporating Content-based Preference and Neighborhood Effect

According to our previous investigation, both content-based preference and neighborhood effect could benefit POI recommendation. In this Section, we exploit both of them for POI recommendation.

Content-based Preference. Once we obtain the ASPECT-MATCHING between a user and a POI, we can modify the predicted preference as

$$\hat{R_{i,j}} = \gamma a_{i,j} + (1-\gamma) U_i L_j^\intercal \tag{4.11}$$

where $\gamma \in [0, 1]$ is a parameter that controls the influence of the content-based preference.

Neighborhood Effect. After incorporating neighborhood features, the predicted preference becomes

$$\hat{R_{i,j}} = \delta NF_{i,j} + (1-\delta) U_i L_j^\intercal \tag{4.12}$$

where $\delta \in [0, 1]$ is a parameter that controls the influence of the neighborhood effect, and $NF_{i,j}$ represents neighborhood features. In this model, we can investigate the effect of exploiting personalized features and property features, independently and jointly.

Content-based Preference and Neighborhood Effect. Accordingly, the predicted preference after incorporating both of them is

$$\hat{R_{i,j}} = \gamma a_{i,j} + \delta NF_{i,j} + (1-\gamma-\delta) U_i L_j^\intercal \tag{4.13}$$

4.6.3 Fusing Geographical Influence

The predicted preference obtained from matrix factorization ignores the geographical influence, which is important in POI recommendation. Hence, we fuse a user's predicted preference for a POI with the geographical preference $G_{i,j}$:

$$P_{i,j} = \hat{R_{i,j}} \cdot G_{i,j} \tag{4.14}$$

The geographical influence models the probability of a user i visits a POI j. We calculate $G_{i,j}$ by the following Equation, which is proposed by Ye et al. [Ye et al., 2011].

$$G_{i,j} = a \cdot (d_{i,j})^b \tag{4.15}$$

where $d_{i,j}$ is the geographical distance between the representative location of user i and POI j; a and b are parameters.

For each user and all of her unvisited POIs, we compute the probability $P_{i,j}$, and select the top-k with the highest probabilities for recommendation.

4.7 Evaluation of POI Recommendation

Evaluation Strategy. Similar to the strategy in Section 4.5, we apply five-fold cross validation for each user. To address the one-class collaborative filtering problem, we sample 10% of unobserved POIs from the training matrix, and treat them as negative samples [Pan and Scholz, 2009].

Evaluation Metric. We define a *discovered* POI of a user as a recommended POI that is actual visited by the user in the test set. Therefore, we evaluate the quality of POI recommendation by the following two metrics:

- Precision@N = No. of discovered POIs / No. of recommended POIs
- Recall@N = No. of discovered POIs / No. of actual visited POIs

Results. We defer the work of comparing our proposed POI recommendation models with state-of-the-art techniques to future work. We are going to compare our models with the following state-of-the-art techniques:

- Aspect and sentiment based approach [Zhang et al., 2015]
- Neighborhood effect based approach [Liu et al., 2014]
- Geographical effect based approach [Ye et al., 2011]

4.8 Related Work

Recommender System Strategies. Recommender system strategies can be broadly divided into two categories: *content filtering* and *collaborative filtering*.

Content filtering approaches build profiles for each user and item to each important nature. For example, a nature of a book could be its author, category, publisher, and so on, and a nature of a user could be her age, education, and any information provided on a certain questionnaire. The profiles allow the system to find an item that best match a user's profile. However, content filtering approaches require a lot of external information, which is difficult to obtain. Therefore, it is less popular than the collaborative filtering approaches.

The common idea of collaborative filtering is to rely on historical behavior to analyze relationships between users and interdependencies between items to infer new user-item associations. Although it is domain free and does not require external information, collaborative filtering approaches suffer the "cold start" problem, where a new user or a new item couldn't be addressed due to the lack of historical data. In this aspect, content filtering approaches are better.

Collaborative filtering include *neighborhood methods* and *latent factor models*. Neighborhood methods can be further divided into *user-based* and *item-based* approaches. Memory-based approaches are centered on computing the relationships between users, or between items. User-based approaches assume that similar users are interested in the same items, and item-based approaches assume that a user tend to be interested in similar items.

Latent factor models aim at characterizing both users and items on a set of factors (e.g., 20 to 100) inferred from observed user-item interaction. The interaction could be a move rating, the visiting frequency from a user to a item, and so on. Such factors can be regarded as an alternative to the aforementioned profiles built from the content filtering approaches. However, they are less interpretable. A factor of a user measures how much that she is interested in the corresponding factor of the item. Some of the most famous latent factor models are based on *matrix factorization* [Koren et al., 2009]. A basic matrix factorization model factorizes both users and items into a joint latent factor space of a certain dimensionality, and the user-item interaction can be obtained by the inner products in that space. Salakhutdinov and Mnih [Salakhutdinov and Mnih, 2008] present a Probabilistic Matrix Factorization (PMF) model, which is the foundation of a lot of recommendation algorithms.

POI Recommendation The POI recommendation approaches can be categorized into five categories: pure check-in data based, geographical influence enhanced, social influence enhanced, temporal influence enhanced, and content information enhanced.

For the pure check-in data based approaches, they treat POIs the same with the traditional items, e.g., movies and books. However, most LBSNs lack the rating data, thus the check-in frequency is regarded as an implicit preference from a user for a POI. Approaches in this category adopt traditional recommendation techniques to the user-POI check-in frequency matrix, without any additional information. Berjani et al. [Berjani and Strufe, 2011] apply regularized matrix factorization to the user-POI check-in frequency matrix. Nevertheless, due to the various information from different data sources, only a few works purely focus on the check-in matrix because it is reported that additional information could improve the performance of POI recommendation.

Geographical influence plays a significant role in the POI recommendation. It assumes that the geographical proximities of POIs is an important factor for users to decide where to visit: users tend to visit POIs that are near to their homes or offices, and they may visit nearby POIs of a POI that they just visited. Ye et al. [Ye et al., 2011] assume that the distance of visited POIs follows a power law distribution, and they propose a a collaborative recommendation algorithm by fusing the geographical influence. Instead of making the power law assumption, Cheng et al. [Cheng et al., 2012] model the geographical influence as a multicenter Gaussian model. That is, users tend to check in around several centers, and check-in POIs around each center follow a Gaussian distribution. Zhang et al. [Zhang and Chow, 2013] argue that the geographical influence on each user should be personalized, and they develop a kernel density estimation approach to personalize the geographical influence on users' check-in behaviors as individual distributions rather than a universal distribution for all users. Their experimental results show that their model outperforms the approaches that are based on power law distribution ([Ye et al., 2011]) or multicenter Gaussian distribution ([Cheng et al., 2012]). In addition, several other researchers build probabilistic graphical models to integrate the geographical influence [Liu et al., 2013a; Kurashima et al., 2013].

The use a Gaussian distribution to represent POIs visited by a user over a sample region. The geographical effect modeled by above works is mainly about the time cost or

monetary cost of travel from a location to another. Differently, we focus on the influence of environmental context (e.g, category, popularity, comments from other users) created by nearby POIs on a POI.

Inspired by the effect of social influence on the traditional recommender systems [Golbeck, 2006; Ma et al., 2008; Jamali and Ester, 2009; Ye et al., 2012], a lot of researchers exploit the social influence to the recommendation algorithms of LBSNs. The basic assumption is that friends would share more common interests of POIs than non-friends. Ye et al. [Ye et al., 2010] propose a friend-based collaborative filtering approach; it only considers the friends of a user instead of all users when applying the user-based collaborative filtering algorithm. Their results show that the approach yields minor improvement over state-of-the-art approaches. Most approaches fuse social influence with geographical influence [Ye et al., 2011; Cheng et al., 2012; Zhang and Chow, 2013]. Gao et al. [Gao et al., 2012b] propose a geo-social correlation model to capture the check-in behavior on LBSNs. They find that social correlations can be leveraged to solve the "cold start" problem. Overall, results of these works show that the improvement of performance on LBSNs after applying social influence is limited, compared with geographical influence.

The assumption underlying temporal influence is that users tend to visit different POIs at different time in a day. For example, a user is more likely to go to restaurants at noon and go to libraries in the evening. Yuan et al. [Yuan et al., 2013b] enhance the user-based collaborative filtering by incorporating the temporal influence: the similarity between users is captured via timed-based check-in vectors. Gao et al. [Gao et al., 2013] propose a matrix factorization based recommendation framework based on two assumptions: (1) a user has personalized check-in preferences at different hours in a day, (2) for a user, the check-in preferences are more similar in consecutive hours than in non-consecutive hours. Both of the approaches find that the temporal influence could improve the performance of POI recommendation.

Due to the rich content information (e.g., reviews, categories of POIs) available on the LBSNs, a lot of researchers enhance their recommendation approaches via doing the content analysis. Liu et al. [Liu and Xiong, 2013] learn topic distributions of users and POIs from textual information to infer the extent to which a user is interested in a POI. They then build a unified framework that considers both interests matching

and POI popularity. However, they ignore the sentiments from a user to a POI and don't differentiate the categories of POIs when learning users' topic distributions. Hu et al. [Hu and Ester, 2013] assume that textual information reflects a user's interests, and propose a probabilistic graphical model to capture the relationship among users' locations, interests, and the function of locations.

Motivated by the importance of sentiment analysis in other applications, researchers start to analyze users' sentiments from their textual information, and develop the recommendation algorithms by incorporating sentiments. Yang et al. [Yang et al., 2013] extend matrix factorization model by exploiting sentiments extracted from users' reviews. Different from our approach, they ignore different aspects covered by a review, and only use a single overall sentiment score to build the model. Similarly, Gao et al. [Gao et al., 2015] extract a single overall sentiment score from each review to scale the importance of a check-in activity. In addition, they construct word-frequency matrix from reviews and descriptions of POIs to represent a user's interests and POI properties. Combined with the three content information sources, they develop a matrix factorization based recommendation framework. Zhang et al. [Zhang et al., 2015] detect sentiments from a user to a POI via a supervised learning framework, and devise a recommendation method that fuses the sentiments with geographical and social influence. Even though they consider several aspects covered by a review to detect sentiments, the final preference modeling from a user to a POI only integrates a single overall sentiment score, which is different from our approach. All of above approaches find that sentiment analysis could provide minor performance improvement for POI recommendation.

There are other works that put an emphasis on the categories of POIs. The intuition is that the category of a POI captures the function of a POI, and a user's visiting behavior at a POI reflects her interests in the corresponding category. Liu et al. [Liu et al., 2013b] learn users' preference transition pattern over categories of POIs, and exploit such pattern to build a matrix factorization based recommendation algorithm. Bao et al. [Bao et al., 2012] models each user's preferences with a weighted category hierarchy, and infers local experts with respect to different categories. Their recommendation algorithm then selects local experts in a user specified geospatial range to best match a user's preferences. Zhang et al. [Zhang and Chow, 2015] weight a POI's popularity by a user's categorical bias to

model the user's preference for the POI. Noulas et al. [Noulas et al., 2012] explicitly define features that capture a user's categorical preferences to predict next check-in place for users.

4.9 Conclusions

To improve the recommendation quality, we investigate two contextual factors: content information and neighborhood effect, and propose a recommendation framework that jointly models both of them. We extract users' aspect-based sentiments toward different POIs to model a user's preference for a POI. We define a set of features to model different properties of the neighborhood effect. We propose a framework that fuses both of them with matrix factorization techniques. Evaluation shows that our method of modeling a user' preference for a POI through the aspect-based sentiment could reduce RMSE of POI recommendation significantly. After investigating different threshold of defining neigbhorhood, we find that 500 m is a better choice than 200 m. In addition, properties features are significantly more effective in modeling neighborhood effect than personalized features. We defer the evaluation of our recommendation models to future work.

Chapter 5

Conclusions

This dissertation investigates three user-interaction problems in social media by exploiting content information and homophily principle. The objective is to overcome data sparsity and enhance our understanding of user interactions.

In Chapter 2, we study the effects of exploiting information from the content layer (sentiment) along with homophily in indicating a user-user interaction. Link prediction refers to inferring potential relationships from a snapshot of a social network. Even though much research has been conducted on this problem, most prior works are based on homophily of data from the social layer [Liben-Nowell and Kleinberg, 2003] or from the geographical layer [Scellato et al., 2011]. The rich data in the content layer has been neglected. Motivated by the easy availability of content, we investigate whether applying homophily to the content information would benefit link prediction.

To the best of our knowledge, only a few researchers exploit content information for link prediction [Romero et al., 2013]. However, users may express different sentiments toward a common semantical interest. Therefore, the first piece of the dissertation is to exploit *sentiment homophily* for link prediction.

We evaluate our approach on a Twitter dataset gathered from U.S. 2012 political campaign. First, based on two users' sentiments toward topics of mutual interest, we define a set of sentiment-based features that quantify the likelihood of them becoming "friends". Our results suggest that sentiment features significantly improve the performance of link prediction in terms of F_1 in both mutual-follow and mention graphs. We further inves-

tigate each predictor and find that Adamic-Adar and Euclidean distance measures are the best. Second, we propose a factor graph model that incorporates a sentiment-based variant of cognitive balance theory. Compared with traditional machine learning techniques, our proposed model is more effective in link prediction when the tie strength is not too weak. In future work, we will evaluate our approach in a friend recommendation framework by exploiting temporal information regarding how links form.

To overcome location sparsity and improve the performance of user-location interactions, Chapter 3 and Chapter 4 investigate two important problems.

Chapter 3 discusses an approach for estimating the locations where the messages originated by exploiting homophily to both the social layer and geographical layers. Since August 2009, Twitter supported per-tweet geo-tagging, i.e., each tweet is associated with a geo-tag. However, geo-tagged tweets are sparse; only about 2% of tweets are geo-tagged [Leetaru et al., 2013]. Location estimation refers to the problem of assigning a geo-tag to a tweet, indicating where the messages originated. Location estimation is important in many applications such as health, urban planning, and advertising.

To overcome the sparsity of available locations, our approach *Percimo* (1) employs communities by applying principle to both the social and geographical layers, (2) relates a user's interests to another user's locations via an LDA-based model that balances a user's personal and communal interests.

We evaluate Percimo via a Twitter geo-tagged dataset collected over two months from two U.S. states. We find that Percimo yields a smaller prediction error than two state-of-the-art approaches: (1) Percimo outperforms a purely content-based technique by reducing the size of candidate sets through communities, (2) Percimo yields a smaller prediction error over an individualized technique by differentiating a user's community interests from personal interests and predicts geo-tags even for users without historical geo-tags. Future work will model the assumption that users could participate in multiple communities.

Lastly, Chapter 4 presents an approach for personalized POI recommendation by exploiting content (aspect-based sentiment) and applying homophily to the geographical layer. Improving the quality of POI recommender systems brings benefits to both users and service providers. To overcome the sparsity of the user-POI check-in matrix and

improve the recommendation quality, we propose a framework that jointly models content and geographical homophily.

We conduct our analysis on the Yelp Challenge Dataset. First, we develop an LDA-based approach to model a user's personalized preference for a POI by exploiting aspect-based sentiment and POI categories. Evaluation demonstrates that our proposed content-based preference modeling approach significantly improves POI recommendation quality in terms of RMSE. Second, we propose features to support different properties of the neighborhood effect. We find that by choosing the threshold 500 m to define a neighborhood, RMSE scores are significantly reduced after applying the neighborhood effect. In addition, features based on neighborhood properties are significantly better than features based on a user's preference toward nearby POIs for the effect of modeling the neighborhood effect. Finally, we propose a unified POI recommendation framework by fusing the content-based preference modeling and the neighborhood effect. We defer the evaluation of recommendation models to future work.

Most of the investigations in this dissertation are based on data from the content, social, and geographical layers. We will investigate exploiting data from the temporal layer to model user interactions in the future.

BIBLIOGRAPHY

Lada A. Adamic and Eytan Adar. Friends and neighbors on the web. *Social Networks*, 25:211–230, 2001.

Gediminas Adomavicius and Alexander Tuzhilin. Context-aware recommender systems. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, pages 335–336, Lausanne, Switzerland, 2008. ACM.

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Languages in Social Media*, pages 30–38, Portland, Oregon, 2011. Association for Computational Linguistics.

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1568–1576, Edinburgh, United Kingdom, 2011. Association for Computational Linguistics.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA), May 2010.

Lars Backstrom and Jure Leskovec. Supervised random walks: Predicting and recommending links in social networks. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, pages 635–644. ACM, 2011.

Lars Backstrom, Eric Sun, and Cameron Marlow. Find me if you can: Improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th International Conference on World Wide Web*, pages 61–70. ACM, 2010.

Akshat Bakliwal, Jennifer Foster, Jennifer van der Puil, Ron O'Brien, Lamia Tounsi, and Mark Hughes. Sentiment analysis of political tweets: Towards an accurate classifier. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 49–58. Association for Computational Linguistics, June 2013.

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

Betim Berjani and Thorsten Strufe. A recommendation system for spots in location-based online social networks. In *Proceedings of the 4th Workshop on Social Network Systems*, pages 4:1–4:6, Salzburg, Austria, 2011. ACM.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, March 2003.

Johan Bollen, Huina Mao, and Alberto Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the International Conference on Weblogs and Social Media*, 2011.

Swarup Chandra, Latifur Khan, and Fahad Bin Muhaya. Estimating Twitter user location using social interactions-a content based approach. In *Proceedings of the IEEE 3rd International Conference on Social Computing*, pages 838–843. IEEE, 2011.

Yan Chen, Jichang Zhao, Xia Hu, Xiaoming Zhang, Zhoujun Li, and Tat-Seng Chua. From interest to function: Location estimation in social media. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*. AAAI Press, 2013.

Chen Cheng, Haiqin Yang, Irwin King, and Michael R. Lyu. Fused matrix factorization with geographical and social influence in location-based social networks. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*. AAAI Press, 2012.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: A content-based approach to geo-locating Twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 759–768, Toronto, ON, Canada, 2010. ACM.

Zhiyuan Cheng, James Caverlee, Kyumin Lee, and Daniel Z. Sui. Exploring millions of footprints in location sharing services. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*. The AAAI Press, 2011.

Kai-Yang Chiang, Nagarajan Natarajan, Ambuj Tewari, and Inderjit S. Dhillon. Exploiting longer cycles for link prediction in signed networks. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 1157–1162, Glasgow, Scotland, UK, 2011. ACM.

Eunjoon Cho, Seth A. Myers, and Jure 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*, pages 1082–1090, San Diego, California, USA, 2011. ACM.

Aaron Clauset, M.E.J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E.*, 70:066111, December 2004.

Aaron Clauset, Cristopher Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453:98–101, May 2008.

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

Sophia Daskalaki, Ioannis Kopanas, and Nikolaos Avouris. Evaluation of classifiers for an uneven class distribution problem. *Applied Artificial Intelligence*, 20:381–417, September 2006.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 449–454, 2006.

Nicholas A. Diakopoulos and David A. Shamma. Characterizing debate performance via aggregated Twitter sentiment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1195–1198. ACM, 2010.

Qiming Diao, Jing Jiang, Feida Zhu, and Ee-Peng Lim. Finding bursty topics from microblogs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 536–544, Jeju Island, Korea, 2012. Association for Computational Linguistics.

Yuxiao Dong, Jie Tang, Sen Wu, Jilei Tian, Nitesh V. Chawla, Jinghai Rao, and Huanhuan Cao. Link prediction and recommendation across heterogeneous social networks. In *Proceedings of the 12th IEEE International Conference on Data Mining*, pages 181–190, 2012.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370, Ann Arbor, Michigan, 2005. Association for Computational Linguistics.

David Flatow, Mor Naaman, Ke Eddie Xie, Yana Volkovich, and Yaron Kanza. On the accuracy of hyper-local geotagging of social media content. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 127–136, Shanghai, China, 2015. ACM.

Foursquare. Foursquare category hierarchy, 2015. URL `https://developer.foursquare.com/categorytree`.

Huiji Gao, Jiliang Tang, and Huan Liu. Exploring social-historical ties on location-based social networks. In *Proceedings of the 7th International Conference on Weblogs and Social Media*. The AAAI Press, 2012a.

Huiji Gao, Jiliang Tang, and Huan Liu. gSCorr: Modeling geo-social correlations for new check-ins on location-based social networks. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 1582–1586. ACM, 2012b.

Huiji Gao, Jiliang Tang, Xia Hu, and Huan Liu. Exploring temporal effects for location recommendation on location-based social networks. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 93–100, Hong Kong, China, 2013. ACM.

Huiji Gao, Jiliang Tang, Xia Hu, and Huan Liu. Content-aware point of interest recommendation on location-based social networks. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*. AAAI Press, 2015.

Jennifer Golbeck. Generating predictive movie recommendations from trust in social networks. In *Proceedings of the 4th International Conference on Trust Management*, pages 93–104. Springer-Verlag, 2006.

Marta C. González, Cesar A. Hidalgo R., and Albert-László Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.

Przemyslaw A. Grabowicz, Luca Maria Aiello, Victor M. Eguiluz, and Alejandro Jaimes. Distinguishing topical and social groups based on common identity and bond theory. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, pages 627–636, Rome, Italy, 2013. ACM.

Mark S. Granovetter. The strength of weak ties. *The American Journal of Sociology*, 78 (6):1360–1380, May 1973.

Mangesh Gupte and Tina Eliassi-Rad. Measuring tie strength in implicit social networks. In *Proceedings of the 3rd Annual ACM Web Science Conference*, pages 109–118. ACM, 2012.

Davide Feltoni Gurini, Fabio Gasparetti, Alessandro Micarelli, and Giuseppe Sansonetti. A sentiment-based approach to Twitter user recommendation. In *Proceedings of the 5th ACM RecSys Workshop on Recommender Systems and the Social Web*. ACM, 2013.

J. M. Hammersley and P. E. Clifford. Markov random fields on finite graphs and lattices. Unpublished manuscript, 1971.

Vasileios Hatzivassiloglou and Janyce M. Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th Conference on Computational Linguistics Volume 1*, pages 299–305. Association for Computational Linguistics, 2000.

Fritz Heider. *The Psychology of Interpersonal Relations*. Wiley, 1958.

Keith Henderson, Tina Eliassi-Rad, Spiros Papadimitriou, and Christos Faloutsos. Hcdf: A hybrid community discovery framework. In *Proceedings of SIAM International Conference on Data Mining*, pages 754–765, Columbus, Ohio, USA, 2010. SIAM.

Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1):5–53, January 2004.

Tuan-Anh Hoang and Ee-Peng Lim. On joint modeling of topical communities and personal interest in microblogs. In *Proceedings of the 6th International Conference on Social Informatics*, pages 1–16, Barcelona, Spain, 2014. Springer.

Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsioutsiouliklis. Discovering geographical topics in the Twitter stream. In *Proceedings of the 21st International Conference on World Wide Web*, pages 769–778, Lyon, France, 2012. ACM.

John Hopcroft, Tiancheng Lou, and Jie Tang. Who will follow you back?: Reciprocal relationship prediction. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 1137–1146. ACM, 2011.

Bo Hu and Martin Ester. Spatial topic modeling in online social media for location recommendation. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 25–32, Hong Kong, China, 2013. ACM.

Mohsen Jamali and Martin 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*, pages 397–406, Paris, France, 2009. ACM.

David Jurgens. That's what friends are for: Inferring location in online social media platforms based on social relationships. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*. The AAAI Press, 2013.

Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. "i'm eating a sandwich in glasgow": Modeling locations with tweets. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, pages 61–68, Glasgow, Scotland, UK, 2011. ACM.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, August 2009.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 03 1951.

Takeshi Kurashima, Tomoharu Iwata, Takahide Hoshide, Noriko Takaya, and Ko Fujimura. Geo topic model: Joint modeling of user's activity area and interests for location recommendation. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining*, pages 375–384. ACM, 2013.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, pages 591–600. ACM, 2010.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann Publishers, 2001.

Kalev Leetaru, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Eric Shook. Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday*, 18 (5), 2013.

Vincent Leroy, B. Barla Cambazoglu, and Francesco Bonchi. Cold start link prediction. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 393–402, Washington, DC, USA, 2010. ACM.

Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 462–470. ACM, 2008.

Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Predicting positive and negative links in online social networks. In *Proceedings of the 19th International Conference on World Wide Web*, pages 641–650, Raleigh, North Carolina, USA, 2010. ACM.

Chenliang Li and Aixin Sun. Fine-grained location extraction from tweets with temporal awareness. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–52, Gold Coast, Queensland, Australia, 2014. ACM.

David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. In *Proceedings of the 12th International Conference on Information and Knowledge Management*, pages 556–559. ACM, 2003.

Ryan N. Lichtenwalter, Jake T. Lussier, and Nitesh V. Chawla. New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 243–252. ACM, 2010.

Michael D. Lieberman, Hanan Samet, and Jagan Sankaranayananan. Geotagging: Using proximity, sibling, and prominence clues to understand comma groups. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, pages 6:1–6:8, Zurich, Switzerland, 2010. ACM.

Bin Liu and Hui Xiong. Point-of-interest recommendation in location based social networks with topic and location awareness. In *Proceedings of the 13th SIAM International Conference on Data Mining*, pages 396–404, Austin, Texas, USA, 2013. SIAM.

Bin Liu, Yanjie Fu, Zijun Yao, and Hui Xiong. Learning geographical preferences for point-of-interest recommendation. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1043–1051, Chicago, Illinois, USA, 2013a. ACM.

Jun S. Liu. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problems. *Journal of the American Statistical Association*, 89(427): 958–966, 1994.

Xin Liu, Yong Liu, Karl Aberer, and Chunyan Miao. Personalized point-of-interest recommendation by mining users' preference transition. In *Proceedings of the 22Nd ACM International Conference on Information and Knowledge Management*, pages 733–738, San Francisco, California, USA, 2013b. ACM.

Yong Liu, Wei Wei, Aixin Sun, and Chunyan Miao. Exploiting geographical neighborhood characteristics for location recommendation. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 739–748, Shanghai, China, 2014. ACM.

Zhengdong Lu, Berkant Savas, Wei Tang, and Inderjit S. Dhillon. Supervised link prediction using multiple sources. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, pages 923–928. IEEE Computer Society, 2010.

Hao Ma, Haixuan Yang, Michael R. Lyu, and Irwin King. Sorec: Social recommendation using probabilistic matrix factorization. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 931–940, Napa Valley, California, USA, 2008. ACM.

Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Where is this tweet from? inferring home locations of Twitter users. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*. The AAAI Press, 2012.

Benjamin Markines, Ciro Cattuto, Filippo Menczer, Dominik Benz, Andreas Hotho, and Gerd Stumme. Evaluating similarity measures for emergent semantics of social tagging. In *Proceedings of the 18th International Conference on World Wide Web*, pages 641–650. ACM, 2009.

Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. HT06, tagging paper, taxonomy, Flickr, academic article, to read. In *Proceedings of the 17th Conference on Hypertext and Hypermedia*, pages 31–40. ACM, 2006.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, August 2001.

Aditya Krishna Menon and Charles Elkan. Link prediction via matrix factorization. In *Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 437–452, Athens, Greece, 2011. Springer-Verlag.

Anna Monreale, Fabio Pinelli, Roberto Trasarti, and Fosca Giannotti. Wherenext: A location predictor on trajectory pattern mining. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 637–646, Paris, France, 2009. ACM.

Kevin P. Murphy, Yair Weiss, and Michael I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, pages 467–475. Morgan Kaufmann Publishers Inc., 1999.

Mark Newman. *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA, 2010.

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

Rong Pan and Marti Scholz. Mind the gaps: Weighting the unknown in large-scale one-class collaborative filtering. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 667–676, Paris, France, 2009. ACM.

M. F. Porter. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137, July 1980.

Deborah A. Prentice, Dale T. Miller, and Jenifer R. Lightdale. Asymmetries in attachments to groups and to their members: Distinguishing between common-identity and common-bond groups. *Personality and Social Psychology Bulletin*, 20(5):484–493, October 1994.

Reid Priedhorsky, Aron Culotta, and Sara Y. Del Valle. Inferring the origin locations of tweets with quantitative confidence. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work*, pages 1523–1536, Baltimore, Maryland, USA, 2014. ACM.

Ross J. Quinlan. Learning with continuous classes. In *Australian Joint Conference on Artificial Intelligence*, pages 343–348, Singapore, 1992. World Scientific.

Daniel M. Romero, Chenhao Tan, and Johan Ugander. On the interplay between social and topical structure. In *Proceedings of the 7th International Conference on Weblogs and Social Media*, 2013.

Mrinmaya Sachan, Danish Contractor, Tanveer Faruquie, and Venkata Subramaniam. Probabilistic model for discovering topic based communities in social networks. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 2349–2352, Glasgow, Scotland, UK, 2011. ACM.

Adam Sadilek, Henry Kautz, and Jeffrey P. Bigham. Finding your friends and following them to where you are. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, pages 723–732, Seattle, Washington, USA, 2012. ACM.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, pages 851–860, Raleigh, North Carolina, USA, 2010. ACM.

Ruslan Salakhutdinov and Andriy Mnih. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*, volume 20, 2008.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, August 1988.

Kai Sassenberg. Common bond and common identity groups on the Internet. *Group Dynamics: Theory, Research, and Practice*, 6(1):27–37, 2002.

Salvatore Scellato, Anastasios Noulas, and Cecilia 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*, pages 1046–1054. ACM, 2011.

Axel Schulz, Aristotelis Hadjakos, Heiko Paulheim, Johannes Nachtwey, and Max Mühlhäuser. A multi-indicator approach for geolocalization of tweets. In *Proceedings of the 8th International Conference on Weblogs and Social Media*, pages 573–582. AAAI Press, 2013.

Parag Singla and Matthew Richardson. Yes, there is a correlation: From social networks to personal behavior on the Web. In *Proceedings of the 17th International Conference on World Wide Web*, pages 655–664. ACM, 2008.

R. W. Sinnott. Virtues of the Haversine. *Sky and Telescope*, 68(2):159, 1984.

Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, September 2010.

Kate Starbird, Leysia Palen, Amanda L. Hughes, and Sarah Vieweg. Chatter on the red: What hazards threat reveals about the social life of microblogged information. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, pages 241–250, Savannah, Georgia, USA, 2010. ACM.

Internet Live Stats. Twitter usage statistics, 2016. URL `http://www.internetlivestats.com/twitter-statistics/`.

Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1397–1405. ACM, 2011.

Wenbin Tang, Honglei Zhuang, and Jie Tang. Learning to infer social ties in large networks. In *Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases*, ECML PKDD, pages 381–397. Springer-Verlag, 2011.

Mike Thelwall. Emotion homophily in social network site messages. *First Monday*, 15 (4), 2010.

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418, February 2011.

W. R. Tobler. A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46:234–240, 1970.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 173–180, Edmonton, Canada, 2003. Association for Computational Linguistics.

Oren Tsur and Ari Rappoport. What's in a hashtag?: Content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, pages 643–652. ACM, 2012.

Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pages 178–185. The AAAI Press, 2010.

Tracy L. Tuten. *Advertising 2.0: Social Media Marketing in a Web 2.0 World*. Greenwood Publishing Group, 2008.

Wikipedia. Preferred walking speed, 2016. URL `https://en.wikipedia.org/wiki/Preferred_walking_speed`.

Dingqi Yang, Daqing Zhang, Zhiyong Yu, and Zhu Wang. A sentiment-enhanced personalized location recommendation system. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 119–128. ACM, 2013.

Shuang-Hong Yang, Bo Long, Alex Smola, Narayanan Sadagopan, Zhaohui Zheng, and Hongyuan Zha. Like like alike: Joint friendship and interest propagation in social

networks. In *Proceedings of the 20th International Conference on World Wide Web*, pages 537–546, Hyderabad, India, 2011. ACM.

Mao Ye, Peifeng Yin, and Wang-Chien Lee. Location recommendation for location-based social networks. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 458–461, San Jose, California, 2010. ACM.

Mao Ye, Peifeng Yin, Wang-Chien Lee, and Dik-Lun 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*, pages 325–334. ACM, 2011.

Mao Ye, Xingjie Liu, and Wang-Chien Lee. Exploring social influence for recommendation: A generative model approach. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 671–680. ACM, 2012.

Yelp. Yelp dataset challenge (round five), 2015a. URL `https://www.yelp.com/dataset_challenge`.

Yelp. Yelp category list, 2015b. URL `https://www.yelp.com/developers/documentation/v2/category_list`.

Zhijun Yin, Liangliang Cao, Quanquan Gu, and Jiawei Han. Latent community topic analysis: Integration of community discovery with topic modeling. *ACM Transactions on Intelligent Systems and Technology*, 3(4):63:1–63:21, September 2012.

Guangchao Yuan, Pradeep K. Murukannaiah, Zhe Zhang, and Munindar P. Singh. Exploiting sentiment homophily for link prediction. In *Proceedings of the 8th ACM Conference on Recommender Systems*, pages 17–24, Foster City, California, USA, 2014. ACM.

Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Nadia Magnenat-Thalmann. Who, where, when and what: Discover spatio-temporal topics for Twitter users. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 605–613, New York, NY, USA, 2013a. ACM.

Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Nadia Magnenat-Thalmann. Time-aware point-of-interest recommendation. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 363–372, Dublin, Ireland, 2013b. ACM.

Haizheng Zhang, C. Lee Giles, Henry C. Foley, and John Yen. Probabilistic community discovery using hierarchical latent gaussian mixture model. In *Proceedings of the 22nd National Conference on Artificial Intelligence*, pages 663–668. AAAI Press, 2007.

Jia-Dong Zhang and Chi-Yin Chow. igslr: Personalized geo-social location recommendation: A kernel density estimation approach. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 334–343. ACM, 2013.

Jia-Dong Zhang and Chi-Yin Chow. Geosoca: Exploiting geographical, social and categorical correlations for point-of-interest recommendations. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 443–452, Santiago, Chile, 2015. ACM.

Jia-Dong Zhang, Chi-Yin Chow, and Yu Zheng. Orec: An opinion-based point-of-interest recommendation framework. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1641–1650, Melbourne, Australia, 2015. ACM.

Elena Zheleva, Hossam Sharara, and Lise Getoor. Co-evolution of social and affiliation networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1007–1016. ACM, 2009.

Vincent W. Zheng, Bin Cao, Yu Zheng, Xing Xie, and Qiang Yang. Collaborative filtering meets mobile recommendation: A user-centered approach. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, pages 236–241. AAAI Press, 2010.