TOWARD GEO-SOCIAL INFORMATION SYSTEMS: METHODS AND ALGORITHMS

A Dissertation

by

ZHIYUAN CHENG

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	James Caverlee
Committee Members,	Robin Murphy
	Frank Shipman
	Daniel Z. Sui
Head of Department,	Nancy Amato

May 2014

Major Subject: Computer Science

ABSTRACT

The widespread adoption of GPS-enabled tagging of social media content via smartphones and social media services (e.g., Facebook, Twitter, Foursquare) uncovers a new window into the spatio-temporal activities of hundreds of millions of people. These “footprints” open new possibilities for understanding how people can organize for societal impact and lay the foundation for new crowd-powered geo-social systems. However, there are key challenges to delivering on this promise: the slow adoption of location sharing, the inherent bias in the users that do share location, imbalanced location granularity, respecting location privacy, among many others. With these challenges in mind, this dissertation aims to develop the framework, algorithms, and methods for a new class of geo-social information systems. The dissertation is structured in two main parts: the first focuses on understanding the capacity of existing footprints; the second demonstrates the potential of new geo-social information systems through two concrete prototypes.

First, we investigate the capacity of using these geo-social footprints to build new geo-social information systems. (i): we propose and evaluate a probabilistic framework for estimating a microblog user’s location based purely on the content of the user’s posts. With the help of a classification component for automatically identifying words in tweets with a strong local geo-scope, the location estimator places 51% of Twitter users within 100 miles of their actual location. (ii): we investigate a set of 22 million check-ins across 220,000 users and report a quantitative assessment of human mobility patterns by analyzing the spatial, temporal, social, and textual aspects associated with these footprints. Concretely, we observe that users follow simple reproducible mobility patterns. (iii): we compare a set of 35 million publicly shared check-ins with a set of over 400 million private query logs recorded by a commercial hotel search engine. Although generated by users with fundamentally different intentions, we find common conclusions may be drawn from both data sources, indicating the viability of publicly shared location information to complement (and replace, in some cases), privately held location information.

Second, we introduce a couple of prototypes of new geo-social information systems that utilize the collective intelligence from the emerging geo-social footprints. Concretely, we propose an activity-driven search system, and a local expert finding system that both take advantage of the collective intelligence. Specifically, we study location-based activity patterns revealed through location sharing services and find that these activity patterns can identify semantically related locations, and help with both unsupervised location clustering, and supervised location categorization with a high confidence. Based on these results, we show how activity-driven semantic organization of locations may be naturally incorporated into location-based web search. In addition, we propose a local expert finding system that identifies top local experts for a topic in a location. Concretely, the system utilizes semantic labels that people label each other, people's locations in current location-based social networks, and can identify top local experts with a high precision. We also observe that the proposed local authority metrics that utilize collective intelligence from expert candidates' core audience (list labelers), significantly improve the performance of local experts finding than the more intuitive way that only considers candidates' locations.

DEDICATION

To Mom, Dad, and Ying.

ACKNOWLEDGEMENTS

Words cannot express all my gratitude to my advisor and my role model, James Caverlee, for his mentorship, advice, guidance, continuous support and encouragement, and above all, his friendship. All I have to say is: choosing him as my advisor was one of the best decisions I have ever made in my life. And after the years of working with, and learning from James, I learned how to do research, and how to become a better man.

I would like to specially thank Robin Murphy, Frank Shipman, Daniel Z. Sui for serving on my dissertation committee, and their advice and support during my Ph.D.

During internships, I was fortunate to have opportunities to work with great mentors, Ken Dallmeyer, Raoul-Sam Daruwala, Thuan Huynh, Wai Gen Yee. And I enjoyed the discussions with my fellow interns, Jonathon Huggins, Oscar Martinez, and Andrew Yates. My collaborators and friends outside Texas A&M have also been helping with this dissertation, including Xiao Liang, Chi Wang, Ke Xu.

Many thanks to my awesome labmates at Infolab. I enjoyed all our office jokes, discussions, and final pushs for paper deadlines. We shared the joys of hearing paper acceptance, and of course the pains of sad faces. You all made my journey full of unforgetable moments. My buddies Kyumin Lee and Krishna Kamath had a tremendous impact on my PhD, and we worked hard together through these years. My labmates Elham Khabiri, Yuan Liang, and Jeff McGee with whom I had wonderful discussions, exchanges of ideas and collaboration on several interesting projects. Said Kashoob, and Chiao-Fang Hsu welcomed me to the lab and helped me understand what it takes to succeed as a graduate student. I enjoyed the dicussions with our on-leave labmates Brian Eoff, and Jeremy Kelley through tweets and emails. I would also like to thank my other labmates Vandana Bachini, Himanshu Barthwal, Cheng Cao, Amir Fayazi, Hancheng Ge, Haokai Lu, Wei Niu.

I am deeply indebted to my parents, Li Cheng, and Aiping He, for always having faith in me in my ups and downs throughout the various stages of my life. Instead of telling me to get straight A's, and who I should be, they encouraged me to be who I want to be. More importantly they have always taught me to be an honest man, and a man that can be helpful to others. My grandparents Jianmin, and Naiying are the constant sources of my inspiration. The rest of my family are equally amazing. I love you all.

My wife and my the best friend Ying, has always been there for me, and endured much because of me. Without the endless love and support from her, none of this work would be possible.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

1. INTRODUCTION

1.1 Motivation

The exponential growth in social media over the past decade has recently been joined by the rise of *location* as a central organizing theme of how users engage with online information services and with each other. Enabled by the widespread adoption of GPS-enabled smartphones, users are now forming a comprehensive *geo-social* overlay of the physical environment of the planet. For example, the Foursquare location sharing service has enabled over 4.5 billion "check-ins" [40], whereby users can link their presence, notes, and photographs to a particular venue. The mobile image sharing service Instagram allows users to selectively attach their latitude-longitude coordinates to each photograph; similar geo-tagged image sharing services are provided by Flickr and a host of other services. And the popular Twitter service sees 500 million Tweets per day, of which around 5 million are tagged with latitude-longitude coordinates. Confirming this trend, a recent Pew Research Center report finds that location is now an increasingly central part of the social media experience:

> Location tagging on social media is up: 30% of social media users now tag their posts with their location. For mobile location services, 74% of smartphone owners get directions or other information based on their current location, and 12% use a geo-social service such as Foursquare to "check in" to locations or share their whereabouts with friends. (Pew Research, Sept 12, 2013) [137]

Compared to proprietary location-based data collected by many entities – e.g., search engine query logs with an associated IP address that can be resolved to a rough location, cell-phone call records that can pinpoint a user to a particular cell tower, and point-of-sale data collected by retailers – these geo-social clues are inherently voluntary and public. As a result, they provide a rich and growing body of geo-location evidence that can potentially support basic scientific inquiry into questions that heretofore were difficult for researchers to study. These difficulties were often due to the proprietary nature of traditional location data, the cost of acquiring new data through small lab-based studies (e.g., due to navigating university IRB protocols, overcoming resistance to personal tracking devices), and the difficulty of sharing such sensitive data with other researchers. Not only do voluntarily shared geo-location cues provide an alternative basis for scientific inquiry, in addition, designers of information management systems (e.g., web search systems, social media discovery, personal information management) can integrate these new public location signals into more robust user models, intelligent "location-aware" services, and so forth. Indeed, we believe that the proliferation of these fine-grained (public) spatio-temporal footprints provides an unprecedented opportunity to gain new insights into:

- The dynamics of human behavior and rhythm/pulsation of social life from local to global levels;
- The dynamics of how ideas spread and how people can organize for societal impact; and
- The development of new geo-social information systems that leverage these global-scale geospatial footprints for real-world impact.

Already, we have witnessed compelling new studies along all three of these dimensions, spanning many research communities – including the data mining and machine learning [7, 21, 30, 33, 71, 98], geographic information systems [27, 45, 70, 110, 122], web search and information retrieval [100, 108], and the emerging computational social science paradigm [53, 60, 75, 64, 69, 103, 112, 126]. For example, the dynamics of fundamental human mobility patterns have been modeled from check-ins mined from two location sharing services – Gowalla and Brightkite – and inherent constraints on these patterns by both geographic and social factors have been discovered [21]. Facebook researchers have provided a comprehensive analysis of the distance between Facebook users, leading to new insights into how social networks are impacted by geography [7]. The LiveHoods [29] project has shown how to identify "living neighborhoods" based on the revealed locations and movements of social media users. And new geo-social information systems have been proposed based on these location cues, including earthquake detection from Twitter information flows [100], a local search system that estimates a user's location utilizing the aggregate signals from the check-ins with real-time contextual information [108], and an event discovery system that organizes spatio-temporal footprints and corresponding media to allow consumers to travel through space and time to experience the world's stories [24].

1.2 Challenges

However, there are key challenges to delivering on the promise of incorporating the collective intelligence from emerging location-based social networks into new geo-social information systems:

Location Granularity: Many users in social media reveal broad, imprecise locations (e.g., at the city or state level), while others provide fine-grained latitude-longitude information. In particular, users are less likely to post precise locations such as street addresses on Twitter and related services. How can these multiple location granularities be integrated to account for uncertainty at different levels?

Bias: Models based on users who do willingly share fine-grained location information will necessarily be biased away from the general population of social media users (and more generally, from the underlying population). How can we model and assess the impact of this bias (and its ultimate impact on applications like local information access or expert finding)?

Sparsity: Personal location-revealing information may be interspersed in an inherently noisy stream of updates reflecting many daily interests (e.g., food, sports, daily chatting with friends). Are there clear location signals embedded in this mix of topics and interests that can be accurately extracted in order to overcome the location sparsity per posted message or per user in online social networks?

Public versus Private Data: Social scientists and geographers have long been interested in modeling the linkages and flows between locations for better understanding a variety of geo-spatial issues that have heavily relied on proprietary data (e.g., query logs and transactions). Given the unprecedented access to the publicly shared data from location-sharing services, can we use the publicly-shared location information via location-sharing services to complement (and replace, in some cases), privately held location information such as that in proprietary query logs?

Privacy: The ubiquity of publicly accessible traces of users via current location-based social networks also causes serious concerns about people's privacy. For example, mining algorithms designed to locate a user based on information "leaked" through social media may be easily mis-used, e.g., for crime and other exploits [34]. Considerable ongoing effort is needed to preserve location privacy while enjoying the benefits of mining location-based social network data.

Lack of Understanding: Last but not the least, even the design space of geo-social information systems is not clearly understood. Do users perceive a difference in ownership of their “location” in scenarios where they explicitly reveal it versus it being inferred from large-scale data-driven approaches (e.g., by applying machine learning approaches)? And how does information access in geo-social information systems differ from traditional web search and friend finding in social networks?

These and related questions lead us to believe that there is a compelling need for new techniques for mining, analyzing, and leveraging geospatial footprints in social media. In the face of these challenges, this dissertation makes a first step toward realizing the potential of these new geo-social systems.

1.3 Contributions

This dissertation research seeks to combine information search and mining capabilities with the collective intelligence of users from online location-based social networks to develop new geo-social information systems. In light of the challenges identified, this dissertation is organized around two fundamental principles: (i) investigating the capacity of geo-social data from publicly available location-based social networks to build a new generation of geo-social information systems; and (ii) demonstrating the potential of new geo-social information systems powered with collective intelligence. Specifically, the dissertation makes the following unique contributions:

- **Overcoming Location Sparsity**: First, in order to tackle the problem of location sparsity, this dissertation is the first to propose the challenge of content-based location estimation in social media. Concretely, we propose and evaluate a probabilistic framework for estimating a user’s location based purely on the content of the user’s posts. The developed location estimator can place 51% of Twitter users within 100 miles of their actual location, relying solely on the public content posted by the user.

- **Investigating Location Sharing Services**: Second, this dissertation presents the first large-scale study of location sharing services through an investigation of 22 million check-ins across 220,000 users. We present the first quantitative assessment of human mobility patterns revealed through these services by analyzing the spatial, temporal, social, and textual aspects associated with these footprints. We find that (i) that locations can be modeled by the activity patterns of people; and (ii) that people follow simple, reproducible patterns.

- **Evaluating Public versus Private Location-Revealing Data**: Third, this dissertation evaluates the capacity of publicly-shared geo-social data to capture real-world flows of people instead of using proprietary data. We compare a set of 35 million publicly shared check-ins with a set of over 400 million private query logs recorded by a commercial hotel search engine. Although generated by users with fundamentally different intentions, we find common conclusions may be drawn from both data sources, indicating the viability of publicly shared location information to complement (and replace, in some cases), privately held location information.

  These first three contributions focus on investigating the capacity of investigating the capacity of the geo-social footprints from emerging location-based social networks to build new generation of geo-social information systems. Based on these observations, we complement these three efforts with our next two contributions focused on demonstrating the potential of new geo-social information systems powered with collective intelligence.

- **Integrating Geo-Social Information into Activity-Driven Local Search**: We propose a prototype *location-based search* system that takes advantage of these new location signals. We study location-based activity patterns revealed through aggregated check-ins from location sharing services and find that these activity patterns can identify semantically related locations, improving both unsupervised location clustering, and supervised location categorization. Based on these results, we show how activity-driven semantic organization of locations may be naturally incorporated into location-based web search.

- **A Geo-spatial Approach to Local Expert Finding**: Finally, we design and evaluate a novel local expert finding system that is built over millions of aggregated location signals. The framework relies on both crowdsourced labels extracted from Twitter as well as expertise propagation through both explicit and implicit social connections. We find high precision and NDCG for the proposed approach in comparison compared to alternative approaches.

1.4 Dissertation Overview

The remainder of this dissertation is organized as follows:

- **Section 2: Overcoming Sparsity: A Content-Driven Approach to Geo-Location** - To tackle the challenge of location sparsity in location-based social networks, we propose and evaluate a probabilistic framework for estimating a microblog user's location based purely on the content of the user's posts.

- **Section 3: Whos, Whats, and Whens of Location Sharing** - Toward better understanding of the properties of people's geo-social footprints from location sharing services, we investigate a set of 22 million check-ins, and report a quantitative assessment of human mobility patterns by analyzing the spatial, temporal, social, and textual aspects associated with these footprints.

- **Section 4: Public Check-ins versus Private Queries: Measuring and Evaluating Spatial Preference** - We investigate the viability of new publicly-available geospatial information to capture spatial preference. Specifically, we compare a set of 35 million publicly shared check-ins voluntarily generated by users of a popular location sharing service with a set of over 400 million private query logs recorded by a commercial hotel search engine.

- **Section 5: Activity-Driven Local Search** - We introduce a location-based search system augmented using activity pattern mined from location-sharing services.

- **Section 6: A Geo-Spatial Approach to Finding Local Experts on Twitter** - We introduce a framework – LocalRank – that utilizes collective intelligence to identify top local experts.

- **Section 7: Conclusions and Future Directions** - We conclude our dissertation contributions, and discuss potential research directions for the results presented here.

2. OVERCOMING SPARSITY: A CONTENT-DRIVEN APPROACH TO GEO-LOCATION*

In this section, we tackle the challenge of location sparsity for location-based social networks. The lack of user adoption of geo-based features per user or per post signals that the promise of microblog services as location-based sensing systems may have only limited reach and impact. Thus, we propose and evaluate a probabilistic framework for estimating a microblog user's location based purely on the content of the user's posts. Our framework can overcome the sparsity of geo-enabled features in these services and bring augmented scope and breadth to emerging location-based personalized information services. Three of the key features of the proposed approach are: (i) its reliance purely on publicly available content; (ii) a classification component for automatically identifying words in posts with a strong local geo-scope; and (iii) a lattice-based neighborhood smoothing model for refining a user's location estimate.

2.1 Introduction

Microblog systems like Twitter contain a huge volume of content, diversified topics, and a wide user bases, which in total provide significant opportunities for mining and exploring the real-time web. Mining this people-centric sensor data promises new personalized information services, including local news summarized from tweets of nearby Twitter users [126], the targeting of regional advertisements, spreading busi-

*Reprinted with permission from "You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users" by Zhiyuan Cheng, James Caverlee, and Kyumin Lee, 2010. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management.* Copyright 2010 by ACM. Reprinted with permission from "A Content-Driven Framework for Geo-locating Microblog Users" by Zhiyuan Cheng, James Caverlee, and Kyumin Lee. *Transactions on Intelligent Systems and Technology*, Volume 4, Issue 1, January 2013. Copyright 2013 by ACM.

ness information to local customers [92], and novel location-based applications (e.g., Twitter-based earthquake detection, which can be faster than through traditional official channels [100]).

Unfortunately, microblog users have been slow to adopt geospatial features. Taking Twitter as an example, as listed in Table 2.1, in a random sample of over 1 million Twitter users, we find that only 21% have listed a user's location as granular as a city name (e.g., Los Angeles, CA); only 5% have listed a location as granular as latitude/longitude coordinates (e.g., 29.3712, -95.2104); the rest are overly general (e.g., California), missing altogether, or nonsensical (e.g., Wonderland). In addition, Twitter began supporting per-tweet geo-tagging in August 2009. Unlike user location (which is a single location associated with a user and listed in each Twitter user's profile), this per-tweet geo-tagging promises extremely fine-tuned Twitter user tracking by associating each tweet with a latitude and longitude. Our sample shows, however, that fewer than 0.42% of all tweets actually use this functionality. Together, the lack of user adoption of geo-based features per user or per post signals that the promise of microblog services as location-based sensing systems may have only limited reach and impact.

To overcome this location sparsity problem, we propose that a reasonable framework to predict a microblog user's location should contain the following features: (i) the proposed framework should be generalizable across social media sites and future human-powered sensing systems; (ii) the framework should be robust in the presence of noise and the sparsity of spatial cues in a microblog user's posts; (iii) the framework should provide accurate and reliable location estimation; and (iv) the prediction framework should be based purely on the publicly available data from the user, with no need for proprietary data from system operators (e.g., backend database) or privacy-sensitive data from users (e.g., IP or user/pass).

With these guidelines in mind, in this manuscript, we propose a framework which is based purely on the content of the user's posts, even in the absence of any other geospatial cues. Our intuition is that a user's posts may encode some location-specific content – either specific place names or certain words or phrases more likely to be associated with certain locations than others (e.g., "howdy" for people from Texas). In this way, we can fill-the-gap for the large portion of microblog users lacking city-level granular location information. By augmenting the massive human-powered sensing capabilities of Twitter and related microblogging services with content-derived location information, this framework can overcome the sparsity of geo-enabled features in these services and bring augmented scope and breadth to emerging location-based personalized information services. This in turn could lead to even broader applications of social media in time-critical situations such as emergency management and tracking the diffusion of infectious diseases.

Effectively geo-locating a microblog user based purely on the content of their posts is a difficult task, however:

- First, microblog status updates are inherently noisy, mixing a variety of daily interests (e.g., food, sports, daily chatting with friends). For example, as shown in Table 2.2, User1 talks about education, C++, conversational topics, and travel. Are there clear location signals embedded in this mix of topics and interests that can be identified for locating a user?

- Second, microblog users often rely on shorthand and non-standard vocabulary for informal communication, meaning that traditional gazetteer terms and proper place names (e.g., Eiffel Tower) may not be present in the content of the posts at all, making the task of determining which terms are location-sensitive non-trivial. As we can see from User2's posts in Table 2.2, User2 relies on

Table 2.1: Categorization of Twitter User's Location Field

Category	Percentage	Example(s)
Coordinates	5%	"29.3712, -95.2104"
City-Level Locations	21%	"Los Angeles, CA", "New York City"
General / Nonsensical / Missing	74%	"California", "Wonderland", NULL

informal language which may cause difficulty in analyzing the user's content.

- Third, even if we could isolate the location-sensitive attributes of a user's posts, a user may have interests that span multiple locations beyond their immediate home location, meaning that the content of their posts may be skewed toward words and phrase more consistent with outside locations. For example, New Yorkers may talk about NBA games in Los Angeles or the earthquake in Haiti. This can also be observed from User1 and User3 in Table 2.2.

- Fourth, a user may have more than one associated location, e.g., due to travel, meaning that content-driven location estimation may have difficulty in precisely identifying a user's location.

With these issues in mind, in this manuscript we propose and evaluate a probabilistic framework for estimating a microblog user's city-level location which satisfies all the requirements we mentioned. Taking only a user's publicly available content as the input data, the framework is generalizable across different microblogging sites, and other on-line social media websites. Experimentally, we select Twitter as an exemplar microblogging service over which to evaluate our framework. The proposed approach relies on three key features: (i) its data input of pure content, without any external data from users or web-based knowledge bases; (ii) a classifier which identifies words in status updates with a local geographic scope; and (iii) a lattice-based neighborhood smoothing model for refining the estimated results. The system

Table 2.2: Examples of Tweets

User	Tweet	Topic	Location Hint
User1	More like this, please. White House science fair: http://bit.ly/9bKI7h	Education	DC
	C++ celebrates 25th anniv of its first commercial release! #TAMU	C++	College Station
	@jelsas I read that as #applausability. I am clapping for your tweet.	Conversation	N/A
	Off to Chicago. Found a Papasito's in concourse E at IAH!	Travel	Chicago / Houston
User2	Shaq dmc. In the place to be. I been doin this here since 93.	Conversation	N/A
	I'm n da apple store. I almost got away a wit dat a new iphone.	Personal	N/A
	Vote for my boy rick fox on dancing wit da stars.	Conversation	N/A
User3	@Peter Dude, were you in San Francisco recently?	Conversation	San Francisco
	Got an email from a guy in Serbia asking for source code.	Personal	Serbia
	Really impressed by fans of the Aggies.	Conversation	TAMU / UC Davis

provides k estimated cities for each user with a descending order of possibility. On average, 51% of randomly sampled microblog users are placed within 100 miles of their actual location (based on an analysis of just 100s of posts). We find that increasing amounts of data (in the form of wider coverage of microblog users and their associated tweets) results in more precise location estimation, giving us confidence in the robustness and continued refinement of the approach.

The rest of this manuscript is organized as follows: Related work is in Section 2.2. Section 2.3 formalizes the problem of predicting a microblog user's geo-location and briefly describes the sampled dataset used in the experiments. In Section 2.4, our estimation algorithm and corresponding refinements are introduced. We present the experimental results in Section 2.5. Finally, conclusions and future work are discussed in Section 5.8.

2.2 Related Work

Studying the geographical scope of online content has attracted attention by researchers in the last decade, including studies of blogs [37, 73], webpages [2], search engine query logs [6], and even web users [83]. Prior work relevant to this manuscript can be categorized roughly into three groups based on the techniques used in geo-locating: content analysis with terms in a gazetteer, content analysis with probabilistic language models, and inference via social relations.

Several studies try to estimate the location of web content utilizing content analysis based on geo-related terms in a specialized external knowledge base (a gazetteer). Amitay et al. [2], Fink et al. [37], and Zong et al. [138] extracted addresses, postal code, and other information listed in a geographical gazetteer from web content to identify the associated geographical scope of web pages and blogs.

Serdyukov et al. [105] generate probabilistic language models based on the tags that photos are labeled with by Flickr users. Based on these models and Bayesian inference, they show how to estimate the location for a photo. In terms of the intention, their method is similar to our work. However, they use a GeoNames database to decide whether a user-submitted tag is a geo-related tag, which can overlook the spatial usefulness of words that may have a strong geo-scope (e.g., earthquake, casino, and so on). Separately, the work of Crandall et al. [28] proposes an approach combining textual and visual features to place images on a map. They have restrictions in their task that their system focuses on which of ten landmarks in a given city is the scope of an image.

In the area of privacy inference, a few researchers have been studying how a user's private information may be inferred through an analysis of the user's social relations. Backstrom et al. [7], Lindamood et al. [74], and Hearthely et al. [51] all share a similar assumption that users related in social networks usually share common attributes. These methods are orthogonal to our effort and could be used to augment the content-based approach taken in this manuscript by identifying common locations among a Twitter user's social network.

Recent work on detecting earthquakes with real-time Twitter data makes use of location information for tracking the flow of information across time and space [100]. Sakaki et al. consider each Twitter user as a sensor and apply Kalman filtering and particle filtering to estimate the center of the bursty earthquake. Their algorithm requires prior knowledge of where and when the earthquake is reported, emphasizing tracking instead of geo-locating users. As a result, this and related methods could benefit from our efforts to assign locations to users for whom we have no location information.

As people care about the privacy issues of real-time microblog systems and location-sharing services, we do note that researchers are working in the opposite direction of trying to protect a user's location information and other sensitive information [9, 57, 42]. Our work could be helpful for researchers in the domain of location-preserving data mining, and raise awareness of the privacy leakages and risks associated with posting location-relevant content to microblogging services.

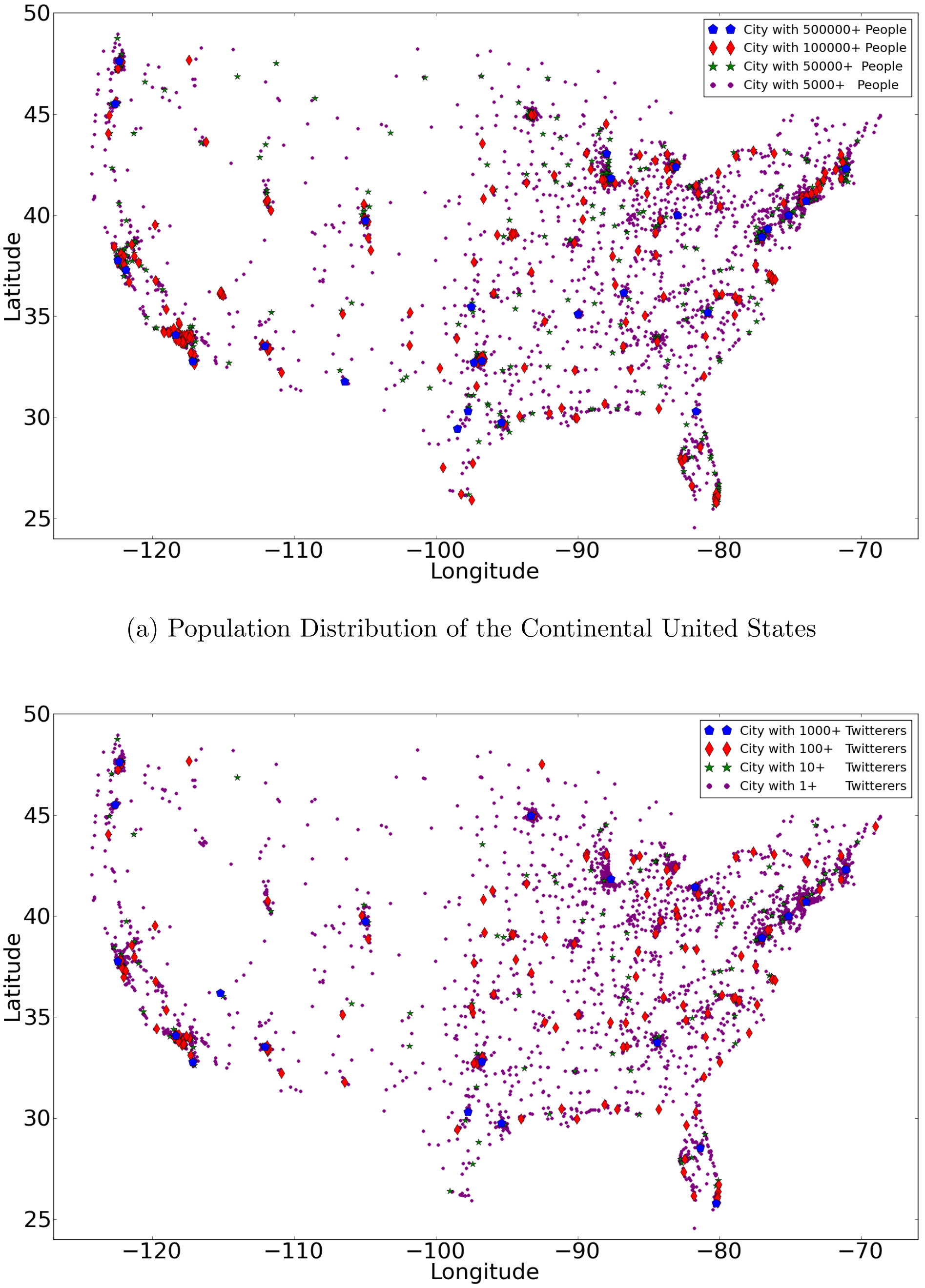
2.3 Preliminaries

In this section, we briefly explain our dataset, formalize the research problem and describe the experimental setup.

2.3.1 Location Sparsity on Twitter

To derive a representative sample of Twitter users, we employed two complementary crawling strategies: crawling through Twitter's public timeline API and crawling by breadth-first search through social edges to crawl each user's friends (following) and followers. The first strategy can be considered as random sampling from active Twitter users (whose tweets are selected for the public timeline), while the second strategy extracts a directed acyclic sub-graph of the whole Twitter social graph, including less active Twitter users. We combine the two strategies to avoid bias in either one. Using the open-source library twitter4j [124] to access Twitter's open API [114] from September 2009 to January 2010, we collected a base dataset of 1,074,375 user profiles and 29,479,600 status updates.

Each user profile includes the capacity to list the user's name, location, a web link, and a brief biography. We find that 72.05% of the profiles collected do list a non-empty location, including locations like "Hollywood, CA", "England", and "UT: 40.708046,-73.789259". However, we find that most of these user-submitted locations are overly general with a wide geographic scope (e.g., California, worldwide), missing altogether, or nonsensical (e.g., Wonderland, "CALI to FORNIA"). Specifically, we examine all locations listed in the 1,074,375 user profiles and find that just 223,418



(a) Population Distribution of the Continental United States

(b) User Distribution of Sampled Twitter Dataset

Figure 2.1: Comparison between the Actual US Population and the Sample Twitter User Population

(21% of the total) list a location as granular as a city name and that only 61,335 (5%) list a location as granular as a latitude/longitude coordinate. This absence of granular location information for the majority of Twitter users (74%) indicates the great potential in estimating or recommending location for a Twitter user.

For the rest of the section, we focus our study of Twitter user location estimation on users within the continental United States. Toward this purpose, we filter all listed locations that have a valid city-level label in the form of "cityName", "cityName, stateName", and "cityName, stateAbbreviation", where we consider all valid cities listed in the Census 2000 U.S. Gazetteer [14] from the U.S. Census Bureau. Even when considering these data forms, there can still be ambiguity for cities listed using just "cityName", e.g., there are three cities named Anderson, four cities named Arlington, and six cities called Madison. For these ambiguous cases, we only consider cities listed in the form "cityName, stateName", and "cityName, stateAbbreviation". After applying this filter, we find that there are 130,689 users (with 4,124,960 status updates), accounting for 12% of all sampled Twitter users. This sample of Twitter users is representative of the actual population of the United States as can be seen in Figure 2.1a, and Figure 2.1b.

2.3.2 Problem Statement

Given the lack of granular location information for Twitter users, our goal is to estimate the location of a user based purely on the content of their tweets. Having a reasonable estimate of a user's location can enable content personalization (e.g., targeting advertisements based on the user's geographical scope, pushing related news stories, etc.), targeted public health web mining (e.g., a Google Flu Trends-like system that analyzes tweets for regional health monitoring), and local emergency detection (e.g., detecting emergencies by monitoring tweets about earthquakes, fires, etc.). By focusing on the content of a user's Twitter stream, such an approach can avoid the need for private user information, IP address, or other sensitive data. With these goals in mind, we focus on city-level location estimation for a Twitter user, where the problem can be formalized as:

Location Estimation Problem: Given a set of tweets $S_{tweets}(u)$ posted by a Twitter user u, estimate a user's probability of being located in city i: $p(i|S_{tweets}(u))$, such that the city with maximum probability $l_{est}(u)$ is the user's actual location $l_{act}(u)$.

As we have noted, location estimation based on tweet content is a difficult and challenging problem. Twitter status updates are inherently noisy, often relying on shorthand and non-standard vocabulary. It is not obvious that there are clear location cues embedded in a user's tweets at all. A user may have interests which span multiple locations and a user may have more than one natural location.

2.3.3 Evaluation Setup and Metrics

Toward developing a content-based user location estimator, we next describe our evaluation setup and introduce four metrics to help us evaluate the quality of a proposed estimator.

2.3.3.1 Test Data

In order to be fair in our evaluation of the quality of location estimation, we build a test set that is separate from the 130,689 users previously identified (and that will be used for building our models for predicting user location). In particular, we extract a set of active users with 1000+ tweets who have listed their location in the form of latitude/longitude coordinates. Since these types of user-submitted locations are typically generated by smartphones, we assume these locations are correct and can be used as ground truth. We filter out spammers, promoters, and other automated-script style Twitter accounts using features derived from Lee et al.'s work

[68] on Twitter bot detection, so that the test set will consist of primarily "regular" Twitter users for whom location estimation would be most valuable. Finally, we arrive at 5,190 test users and more than 5 million of their tweets. These test users are distributed across the continental United States similar to the distributions seen in Figure 2.1a, and Figure 2.1b.

2.3.3.2 Metrics

To evaluate the quality of a location estimator, we compare the estimated location of a user versus the actual city location (which we know based on the city corresponding to their latitude/longitude coordinates). The first metric we consider is the **Error Distance** which quantifies the distance in miles between the actual location of the user $l_{act}(u)$ and the estimated location $l_{est}(u)$. The **Error Distance** for user u is defined as:

$$ErrDist(u) = d(l_{act}(u), l_{est}(u))$$

To evaluate the overall performance of a content-based user location estimator, we further define the **Average Error Distance** across all test users U:

$$AvgErrDist(U) = \frac{\sum_{u \in U} ErrDist(u)}{|U|}$$

A low **Average Error Distance** means that the system can geo-locate users close to their real location on average, but it does not give strong insight into the distribution of location estimation errors. Hence, the next metric – **Accuracy** – considers the percentage of users with their error distance categorized in the range of 0-100 miles:

$$Accuracy(U) = \frac{|\{u|u \in U \wedge ErrDist(u) \leq 100\}|}{|U|}$$

Further, since the location estimator predicts k cities for each user in decreasing order of confidence, we define the **Accuracy with K Estimations (Accuracy@k)** which applies the same Accuracy metric, but over the city in the top-k with the least error distance to the actual location. In this way, the metric shows the capacity of an estimator to identify a good candidate city, even if the first prediction is in error.

2.4 Content-Based Location Estimation: Overview and Approach

In this section, we begin with an overview of our baseline approach for content-based location estimation and then present two key optimizations for improving and refining the quality of location estimates.

2.4.1 Baseline Location Estimation

First, we can directly observe the actual distribution across cities for each word in the sampled dataset. Based on maximum likelihood estimation, the probabilistic distribution over cities for word w can be formalized as $p(i|w)$ which identifies for each word w the likelihood that it was issued by a user located in city i. For example, for the word "rockets", we can see its city distribution in Figure 2.2 based on the tweets in the sampled dataset (with a large peak near Houston, home of NASA and the NBA basketball team Rockets).

Of course users from cities other than Houston may tweet the word "rockets", so reliance on a single word or a single tweet will necessarily reveal very little information about the true location of a user. By aggregating across all words in tweets posted by a particular user, however, our intuition is that the location of the user will become clear. Given the set of words $S_{words}(u)$ extracted from a user's tweets $S_{tweets}(u)$, we propose to estimate the probability of the user being located in city i as:

$$p(i|S_{words}(u)) = \sum_{w \in S_{words}(u)} p(i|w) * p(w)$$

where we use $p(w)$ to denote the probability of the word w in the whole dataset. Letting $count(w)$ be the number of occurrences of the word w, and t be the total number of tokens in the corpus, we replace $p(w)$ with $\frac{count(w)}{t}$ in calculating the value of $p(w)$. Such an approach will produce a per-user city probability across all cities. The city with the highest probability can be taken as the user's estimated location. This location estimator is formalized in Algorithm 1.

Algorithm 1 Content-Based User Location Estimation

Input:
tweets: List of n tweets from a Twitter user u
cityList: Cities in continental US with 5k+ people
distributions: Probabilistic distributions for words
k: Number of estimations for each user
Output:
estimatedCities: Top K estimations

```
$words = preProcess(tweets)$
for $city$ in $cityList$ do
    $prob[city] \leftarrow 0$
    for $word$ in $words$ do
        $prob[city] += distributions[word][city] * word.count$
    end for
end for
$estimatedCities = sort(prob, cityList, k)$
return $estimatedCities$
```

2.4.2 Initial Results

Using this baseline approach, we estimated the location of all users in our test set using per-city word distributions estimated from the 130,689 users shown in

Figure 2.1b. For each user, we parsed their location and status updates (4,124,960 in all). In parsing the tweets, we eliminate all occurrences of a standard list of 319 stop words, as well as screen names (which start with @), hyperlinks, and punctuation in the tweets. Instead of using stemming, we use the Jaccard Coefficient to check whether a newly encountered word is a variation of a previously encountered word. The Jaccard Coefficient is particularly helpful in handling informal content like in tweets, e.g., by treating “awesoome” and “awesooome” as the word “awesome”. In generating the word distributions, we only consider words that occur at least 50 times in order to build comparatively accurate models. Thus, 25,987 per-city word distributions are generated from a base set of 481,209 distinct words.

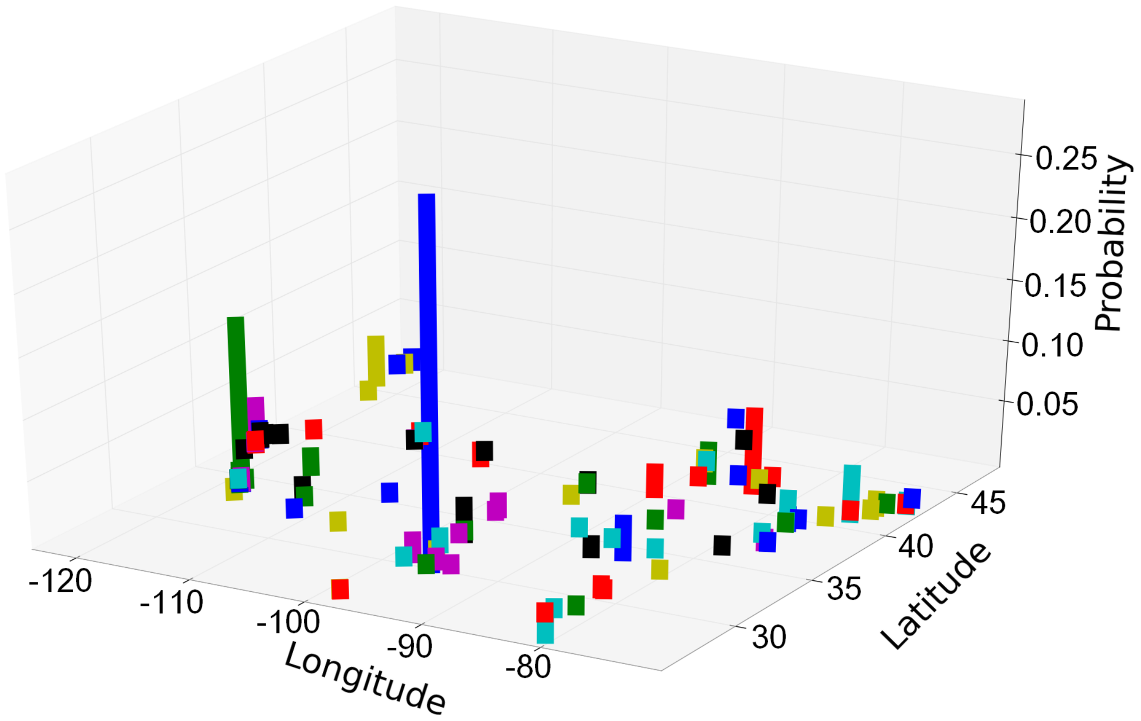


Figure 2.2: City Estimates for the Word “Rockets”

Disappointingly, only 10.12% of the 5,119 users in the test set are geo-located within 100 miles to their real locations and the **AvgErrDist** is 1,773 miles, meaning that such a baseline content-based location estimator provides little value. On inspection, we discovered two key problems: (i) most words are distributed consistently with the population across different cities, meaning that most words provide very little power at distinguishing the location of a user; and (ii) most cities, especially with a small population, have a sparse set of words in their tweets, meaning that the per-city word distributions for these cities are under-specified leading to large estimation errors.

In the rest of this section, we address these two problems in turn in hopes of developing a more valuable and refined location estimator. Concretely, we pursue two directions:

- *Identifying Local Words in Tweets:* Is there a subset of words which have a more compact geographical scope compared to other words in the dataset? And can these "local" words be discovered from the content of tweets? By removing noise words and non-local words, we may be able to isolate words that can distinguish users located in one city versus another.

- *Overcoming Tweet Sparsity:* In what way can we overcome the location sparsity of words in tweets? By exploring approaches for smoothing the distributions of words, can we improve the quality of user location estimation by assigning non-zero probability for words to be issued from cities in which we have no word observations?

2.4.3 Identifying Local Words in Tweets

Our first challenge is to filter the set of words considered by the location estimation algorithm (Algorithm 1) to consider primarily words that are essentially "local".

By considering all words in the location estimator, we saw how the performance suffers due to the inclusion of noise words that do not convey a strong sense of location (e.g., “august”, “peace”, “world”). By observation and intuition, some words or phrases have a more compact geographical scope. For example, “howdy” which is a typical greeting word in Texas may give the estimator a hint that the user is in or near Texas.

Toward the goal of improving user location estimation, we characterize the task of identifying local words as a decision problem. Given a word, we must decide if it is local or non-local. Since tweets are essentially informal communication, we find that relying on formally defined location names in a gazetteer is neither scalable nor provides sufficient coverage. That is, Twitter's 140 character length restriction means that users may not write the full address or location name (e.g., “t-center” instead of “Houston Toyota Center”, home of the NBA Rockets team. Concretely, we propose to determine local words using a model-driven approach based on the observed geographical distribution of the words in tweets.

2.4.3.1 Determining Spatial Focus and Dispersion

Intuitively, a local word is one with a high local focus and a fast dispersion, that is it is very frequent at some central point (like say in Houston) and then drops off in use rapidly as we move away from the central point. Non-local words, on the other hand, may have many multiple central points with no clear dispersion (e.g., words like basketball). How do we assess the spatial focus and dispersion of words in tweets?

Recently Backstrom et al. introduced a model of spatial variation for analyzing the geographic distribution of terms in search engine query logs [6]. The authors propose a generative probabilistic model in which each query term has a geographic focus on a map (based on an analysis of the IP-address-derived locations of users issuing the query term). Around this center, the frequency shrinks as the distance from the center increases. Two parameters are assigned for each model, a constant C which identifies the frequency in the center, and an exponent α which controls the speed of how fast the frequency falls as the point goes further away from the center. The formula for the model is $Cd^{-\alpha}$ which means that the probability of the query issued from a place with a distance d from the center is approximately $Cd^{-\alpha}$. In the model, a larger α identifies a more compact geo-scope of a word, while a smaller α displays a more global popular distribution.

In the context of tweets, we can similarly determine the focus (C) and dispersion (α) for each tweet word by deriving the optimal parameters that fit the observed data. These parameters C and α are strong criteria for assessing a word's focus and dispersion, and hence, determining whether a word is local or not. For a word w, given a center, the central frequency C, and the exponent α, we compute the maximum-likelihood value like so: for each city, suppose all users tweet the word w from the city a total of n times, then we multiply the overall probability by $(Cd_i^{-\alpha})^n$; if no users in the city tweet the word w, we multiply the overall probability by $1 - Cd_i^{-\alpha}$. In the formula, d_i is the distance between city i and the center of word w. We add logarithms of probabilities instead of multiplying probabilities in order to avoid underflow. For example, let S be the set of occurrences for word w (indexed by cities which issued the word w), and let d_i be the distance between a city i and the model's center. Then:

$$f(C, \alpha) = \sum_{i \in S} \log Cd_i^{-\alpha} + \sum_{i \notin S} \log \left(1 - Cd_i^{-\alpha}\right)$$

is the likelihood value for the given center, C and α. Backstrom et al. also prove

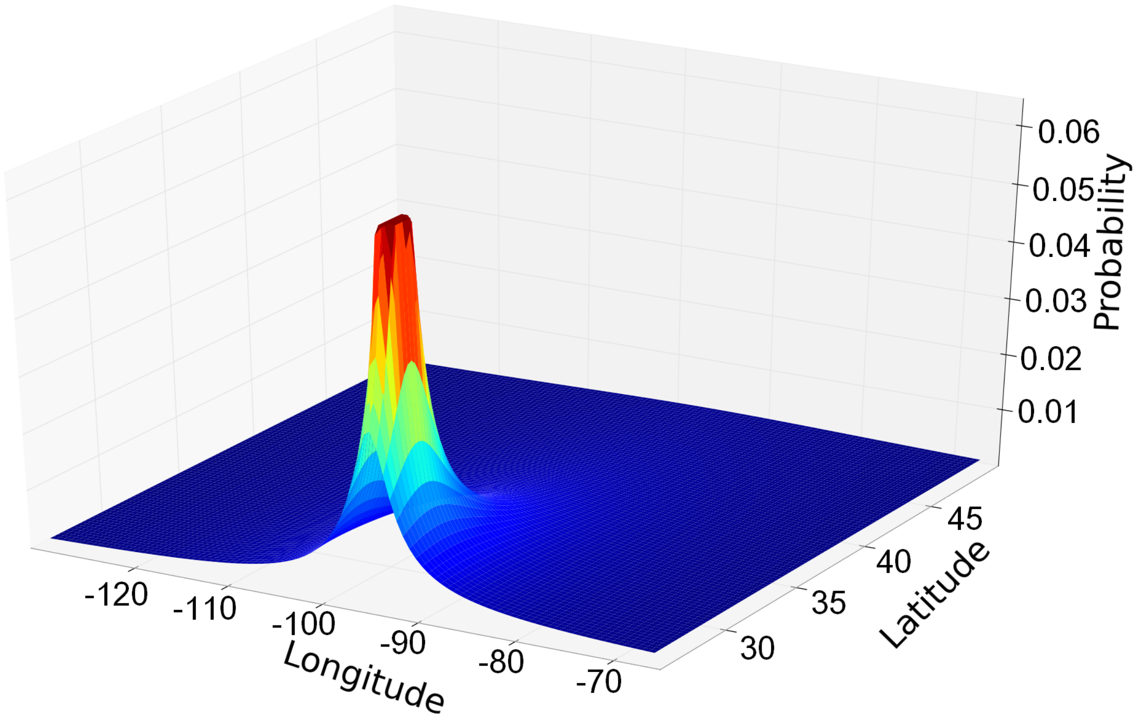


Figure 2.3: Optimized Model for the Word "Rockets"

that $f(C, \alpha)$ has exactly one local maximum over its parameter space which means that when a center is chosen, we can iterate C and α to find the largest $f(C, \alpha)$ value (and hence, the optimized C and α). Instead of using a brute-force algorithm to find the optimized set of parameters, we divide the map of the continental United States into lattices with a size of two by two square degrees. For the center in each lattice, we use golden section search [93] to find the optimized central frequency and the shrinking factor α. Then we zoom into the lattice which has the largest likelihood value, and use a finer-grained mesh on the area around the best chosen center. We repeat this zoom-and-optimize procedure to identify the optimal C, and α. Note that the implementation with golden section search can generate an optimal model for a word within a minute on a single modern machine and is scalable to handle web-scale data. To illustrate, Figure 2.3 shows the optimized model for the word

Table 2.3: Example Local Words

Word	Latitude	Longitude	C_0	α
automobile	40.2	-85.4	0.5018	1.8874
casino	36.2	-115.24	0.9999	1.5603
tortilla	27.9	-102.2	0.0115	1.0350
canyon	36.52	-111.32	0.2053	1.3696
redsox	42.28	-69.72	0.1387	1.4516

“rockets” centered around Houston.

2.4.3.2 Training and Evaluating The Model

Given the model parameters C (focus) and α (dispersion) for every word, we could directly label as local words all tweet words with a sufficiently high focus and fast dispersion by considering some arbitrary thresholds. However, we find that such a direct application may lead to many errors (and ultimately poor user location estimation). For example, some models may lack sufficient supporting data resulting in a clearly incorrect geographic scope. Hence, we augment our model of local words with coordinates of the geo-center, since the geographical centers of local words should be located in the continental United States, and the count of the word occurrences, since a higher number of occurrences of a word will give us more confidence in the accuracy of the generated model of the word.

Using these features, we train a local word classifier using the Weka toolkit [120] – which implements several standard classification algorithms like Naive Bayes, SVM, AdaBoost, etc. – over a hand-labeled set of standard English words taken from the 3esl dictionary [5]. Of the 19,178 words in the core dictionary, 11,004 occur in the sampled Twitter dataset. Using 10-fold cross-validation and the SimpleCart classifier, we find that the classifier has a *Precision* of 98.8% and *Recall* and *F-Measure* both as 98.8%, indicating that the quality of local word prediction is good.

After learning the classification model over these known English words, we apply the classifier to the rest of the 14,983 tweet words (many of which are non-standard words and not in any dictionary), resulting in 3,183 words being classified as local words.

To illustrate the geographical scope of the local words discovered by the classifier, five local word models are listed in Table 2.3. The word “automobile” is located around two hundred miles south of Detroit which is the traditional auto manufacturing center of the US. The word “casino” is located in the center of Las Vegas, two miles east of the North Las Vegas Airport. “tortilla” is centered a hundred miles south of the border between Texas and Mexico. The word “canyon” is located almost at the center of the Grand Canyon. The center for the word “redsox” is located 50 miles east of Boston, home of the baseball team.

In order to visualize the geographical centers of the local favored words, a few examples are shown on the map of the continental United States in Figure 2.4. Based on these and the other discovered local words, we will evaluate if and how user location estimation improves in the experimental study in Section 2.5.

2.4.4 Overcoming Tweet Sparsity

The second challenge for improving our content-based user location estimator is to overcome the sparsity of words across locations in our sampled Twitter dataset. Due to this sparseness, there are a large number of “tiny” word distributions (i.e., words issued from only a few cities) The problem is even more severe when considering cities with a small population. As an example, consider the distribution for the word “rockets” over the map of the continental United States displayed in Figure 2.2. We notice that for a specific word, the probability for the word to be issued in a city can be zero since there are no tweets including the word in our sampled dataset. In order

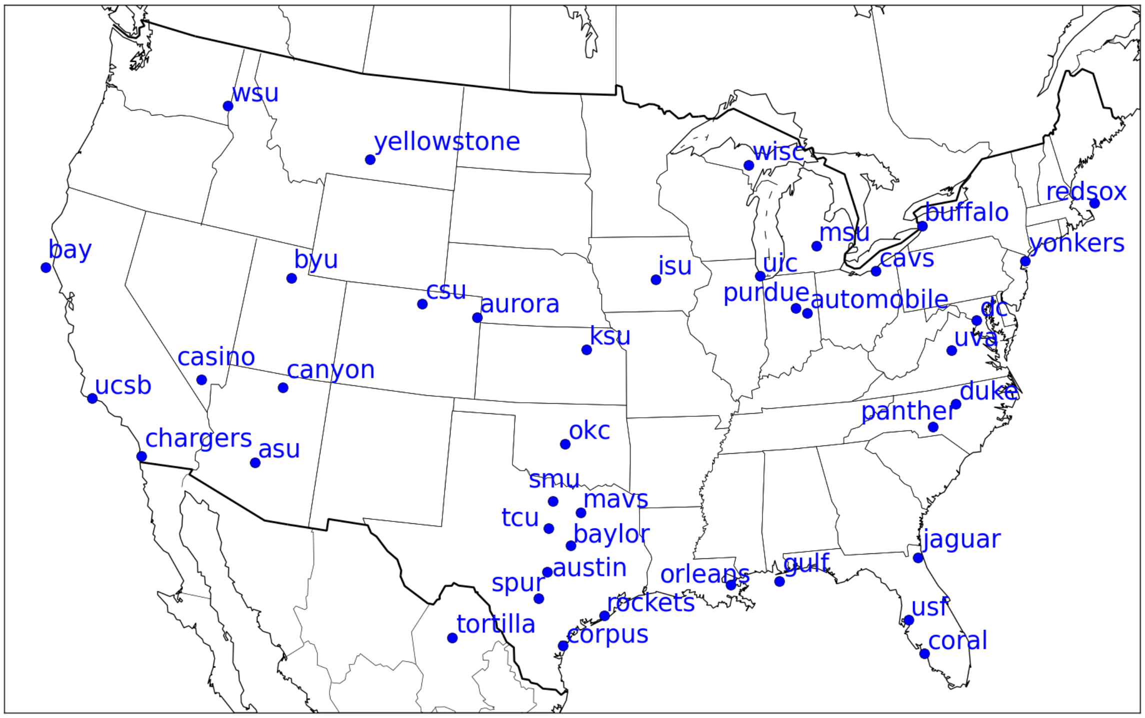


Figure 2.4: Geographical Centers of Local Words Discovered in Sampled Twitter Dataset

to handle this sparsity, we consider three approaches for smoothing the probability distributions: Laplace smoothing, data-driven geographic smoothing, and model-driven smoothing.

2.4.4.1 Laplace Smoothing

A simple method of smoothing the per-city word distributions is Laplace smoothing (add-one smoothing) which is defined as:

$$p(i|w) = \frac{1 + count(w, i)}{V + N(w)}$$

where $count(w, i)$ denotes the term count of word w in city i; V stands for the size of the vocabulary and N(w) stands for the total count of w in all the cities. Briefly speaking, Laplace smoothing assumes every seen or unseen city issued word w once more than it did in the dataset.

Although simple to implement, Laplace smoothing does not take the geographic distribution of a word into consideration. That is, a city near Houston with zero occurrences of the word "rockets" is treated the same as a city far from Houston with zero occurrences. Intuitively, the peak for "rockets" in Houston (recall Figure 2.2) should impact the probability mass at nearby cities.

2.4.4.2 Data-Driven Geographic Smoothing

To take this geographic nearness into consideration, we consider two techniques for smoothing the per-city word distributions by considering neighbors of a city at different granularities. In the first case, we smooth the distribution by considering the overall prevalence of a word within a state; in the second, we consider a lattice-based neighborhood approach for smoothing at a more refined city-level scale.

State-Level Smoothing: For state-level smoothing, we aggregate the probabilities of a word w in the cities in a specific state s (e.g., Texas), and consider the average of the summation as the probability of the word w occurring in the state. Letting S_c denote the set of cities in the state s, the state probability can be formulated as:

$$p_s(s|w) = \frac{\sum_{i \in S_c} p(i|w)}{|S_c|}$$

Furthermore, the probability of the word w to be located in city i can be a combination of the city probability and the state probability:

$$p'(i|w) = \lambda * p(i|w) + (1 - \lambda) * p_s(s|w)$$

where i stands for a city in the state s, and $1-\lambda$ is the amount of smoothing. Thus, a small value of λ indicates a large amount of state-level smoothing.

Lattice-based Neighborhood Smoothing: Naturally, state-level smoothing is a fairly coarse technique for smoothing word probabilities. For some words, the region of a state exaggerates the real geographical scope of a word; meanwhile, the impact of a word issued from a city may have higher influence over its neighborhood in another state than the influence over a distant place in the same state. With this assumption, we apply lattice-based neighborhood smoothing.

Firstly, we divide the map of the continental United States into lattices of 1 x 1 square degrees. Letting w denote a specific word, lat a lattice, and S_c be the set of cities in lat, the per-lattice probability of a word w can be formalized as:

$$p(lat|w) = \sum_{i \in S_c} p(i|w)$$

In addition, we consider lattices around (the nearest lattice in all eight directions) lat as the neighbors of the lattice lat. Introducing μ as the parameter of neighborhood smoothing, the lattice probability is updated as:

$$p'(lat|w) = \mu * p(lat|w) + (1.0 - \mu) * \sum_{lat_i \in S_{neighbors}} p(lat_i|w)$$

In order to utilize the smoothed lattice-based probability, another parameter λ is introduced to aggregate the real probability of w issued from the city i, and the probability of the smoothed lattice probability. Finally the lattice-based per-city word probability can be formalized as:

$$p'(i|w) = \lambda * p(i|w) + (1.0 - \lambda) * p'(lat|w)$$

where i is a city within the lattice lat.

2.4.4.3 Model-Based Smoothing

The final approach to smoothing takes into account the word models developed in the previous section for identifying C and α. Applying this model directly, where each word is distributed according to $Cd^{-\alpha}$, we can estimate a per-city word distribution as: $p'(i|w) = C(w)d_i^{-\alpha(w)}$ where $C(w)$ and $\alpha(w)$ are taken to be the optimized parameters derived from the real data distribution of words across cities. This model-based smoothing ignores local perturbations in the observed word frequencies, in favor of a more elegant word model (recall Figure 2.3). Compared to the data-driven geographic-based smoothing, model-based smoothing has the advantage of "compactness", by encoding each word's distribution according to just two parameters and a center, without the need for the actual city word frequencies.

2.4.4.4 Wave-Like Smoothing:

The term-localizing component works well for terms which have exactly one geographical center. However, some of the words cannot be simply represented by a single peak. Let us still take the word "Rockets" as an example: Rockets is the name of the NBA team in Houston, as well as a term frequently used in NASA which is also located in Houston. Thus people tweet the word "Rockets" the most frequently in the greater Houston area, but there are also "Rockets" associated with the University of Toledo in Ohio and with particular events (like the mysterious rocket launch off the coast of California in 2010).

To handle this multi-peak issue, we can extend the one-peak spatial model to a multiple peaks version. For each word, we generate a peak at each city where the word is issued. In addition, each peak at a city becomes a radioactive source, emitting wave-like impacts towards other cities over the map. The impacts from

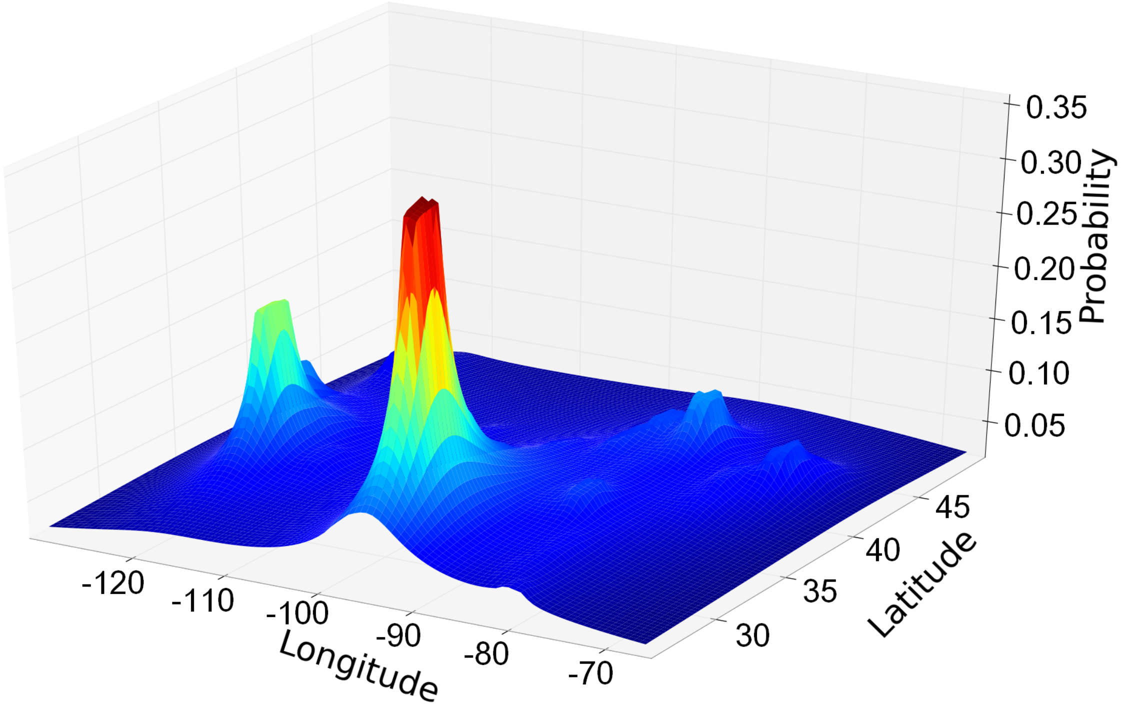


Figure 2.5: Wave-Like Smoothing for Word "Rockets"

each peak (i.e., source) decreases exponentially as the distance from the location of the peak increases. Thus, the probability distribution for a word becomes an interwoven overlapping of thousands of one-peak distributions. We visualize the wave-like distribution for the word "Rockets" in Figure 2.5, and at least three relative high peaks can be identified. With this *wave-like smoothing*, the probability of a word w issued from city i can be formalized as:

$$p(i|w) = \sum_{j \in S_c} \begin{cases} p(j|w) * (d(i,j) - r_j + 1)^{-\alpha(w)} & d(i,j) \geq r_j \\ p(j|w) & d(i,j) < r_j \end{cases}$$

where $p(j|w)$ denotes the estimated probability of word w issued from city j; $d(i,j)$ is the euclidean distance between city i and city j; r_j is the radius of the city j; and $\alpha(w)$ is the shrinking parameter of word w indicating how fast the impacting probability of the word w shrinks down when distance from the center increases. With the equation above, we go through all the cities in the set of large cities S_c and sum up the impacts from each city. Locations inside the area of each source city will have the same probability as the city's $p(j|w)$, and as the distance from the source increases, the probability decreases exponentially. As a consequence, with the combinations of all the local words and all the cities, a highly overlapped probabilistic distribution is generated.

2.4.5 Social Refinement

So far, we explored predicting an individual Twitter user's geolocation based on her tweets alone. A natural hypothesis would be: given an un-located user's tweets and a few of her un-located friends and their tweets, can we improve the performance of predicting the user's location by incorporating evidence from these social ties? Thus, in this section we explore the possibility of geolocating a user utilizing aggregated results from social relations. The assumption of our method is that users have more local friends than distant friends as researchers have previously observed in [7]. The hope is that aggregates of location predictions from a user's social ties will provide additional evidence for refining the user's predicted geolocation.

For each user u, we have a collection of the user's latest tweets $S_{tweets}(u)$, a list of the user's n friends $list_{friends}(u) = \{f_j | 1 \leq j \leq n\}$, and a collection of tweets $S_{tweets}(f_j)$ from each friend f_j. Determining the appropriate choice of friends and the number of friends to consider is something we can study experimentally.

Given the setup, the *social refinement algorithm* for content-driven location estimation is:

- Firstly, we apply the baseline content-driven algorithm to predict the location for each friend f_j of the user u's. Concretely, for each city i, we estimate a probability $p(i|S_{tweets}(f_j))$ for the friend f_j to be located in city i based on her tweets $S_{tweets}(f_j)$.

- Secondly, for each city i, we get an average probability for user u to be located in the city i based the probabilities estimated from her friends' tweets, formalized as $p(i|S_{tweets}(list_{friends}(u)) = \frac{\sum\limits_{f_j \in list_{friends}(u)} p(i|S_{tweets}(f_j))}{|list_{friends}(u)|}$.

- Thirdly, for each city i, we predict the probability for user u to be located in city i based on user u's tweets $S_{tweets}(u)$: $p(i|S_{tweets}(u))$.

- Fourthly, for each city i, the social inferred probability for user u to be located in city i is: $p(i|S_{tweets}(u), S_{tweets}(list_{friends}(u)) = \alpha * p(i|S_{tweets}(list_{friends}(u)) + (1-\alpha)*p(i|S_{tweets}(u))$, where α is a pre-defined weight for predicted probability from social relations.

- Finally, according to the descending order of $p(i|S_{tweets}(u), S_{tweets}(list_{friends}(u))$ we rank the cities, and consider the city with the highest probability as the predicted location for user u.

In this way, the content-driven location estimation algorithm may be enhanced by incorporating the social ties of the underlying social network.

2.5 Experimental Results

In this section, we detail an experimental study of location estimation with local tweet identification and smoothing. The goal of the experiments is to understand: (i) if the classification of words based on their spatial distribution significantly helps improve the performance of location estimation by filtering out non-local words; (ii) how the different smoothing techniques help overcome the problem of data sparseness; (iii) how the amount of information available about a particular user (via tweets) impacts the quality of estimation; and (iv) what impact social refinement has on content-driven location estimation.

Table 2.4: Impact of Refinements on User Location Estimation

Method	ACC	AvgErrDist (Miles)	ACC@2	ACC@3	ACC@5
Baseline	0.101	1773.146	0.375	0.425	0.476
+ Local Filtering (LF)	0.498	539.191	0.619	0.682	0.781
+ LF + Laplace	0.480	587.551	0.593	0.647	0.745
+ LF + State-Level	0.502	551.436	0.617	0.687	0.783
+ LF + Neighborhood	**0.510**	**535.564**	**0.624**	**0.694**	**0.788**
+ LF + Model-based	0.250	719.238	0.352	0.415	0.486
+ LF + Wave-Like	0.507	545.500	0.521	0.530	0.539

2.5.1 Location Estimation: Impact of Refinements

Recall that in our initial application of the baseline location estimator, we found that only 10.12% of the 5,119 users in the test set could be geo-located within 100 miles of their actual locations and that the AvgErrDist across all 5,119 users was 1,773 miles. To test the impact of the two refinements – local word identification and smoothing – we update Algorithm 1 to filter out all non-local words and to update the per-city word probabilities with the smoothing approaches described in the previous section.

For each user u in the test set, the system estimates k (10 in the experiments) possible cities in descending order of confidence. Table 2.4 reports the Accuracy, Average Error Distance, and Accuracy@k for the original baseline user location estimation approach (*Baseline*), an approach that augments the baseline with local word filtering but no smoothing (*+ Local Filtering*), and then five approaches that augment local word filtering with smoothing – *LF+Laplace*, *LF+State-level*, *LF+Neighborhood*, *LF+Model-based*, and *LF+Wave-Like*. Recall that Accuracy measures the fraction of users whose locations have been estimated to within 100 miles of their actual location.

First, note the strong positive impact of local word filtering. With local word filtering alone, we reach an Accuracy of 0.498 which is almost five times as high as the Accuracy we get with the baseline approach that uses all words in the sampled Twitter dataset. The gap indicates the strength of the noise introduced by non-local words, which significantly affects the quality of user location estimation. Also consider that this result means that nearly 50% of the users in our test set can be placed in their actual city purely based on an analysis of the content of their tweets. Across all users in the test set, filtering local words reduces the Average Error Distance from 1,773 miles to 539 miles. While this result is encouraging, it also shows that there are large estimation errors for many of our test users in contrast to the 50% we can place within 100 miles of their actual location. Our hypothesis is that some users are inherently difficult to locate based on their tweets. For example, some users may intentionally misrepresent their home location, say by a New Yorker listing a location in Iran as part of sympathy for the recent Green movement. Other users may tweet purely about global topics and not reveal any latent local biases in their choice of words. In our continuing work, we are examining these large error cases.

Continuing our examination of Table 2.4, we also observe the positive impact of smoothing. Though less strong than local word filtering, we see that Laplace, State-level, Neighborhood, and Wave-Like smoothing result in better user location estimation than either the baseline or the baseline plus local word filtering approach. As we had surmised, the Neighborhood smoothing provides the best overall results, placing 51% of users within 100 miles of their actual location, with an Average Error

Distance over all users of 535 miles.

Comparing State-level smoothing to Neighborhood smoothing, we find similar results with respect to the baseline, but slightly better results for the Neighborhood approach. We attribute the slightly worse performance of state-level smoothing to the regional errors introduced by smoothing toward the entire state instead of a local region. For example, state-level smoothing will favor the impact of words emitted by a city that is distant but within the same state relative to a words emitted by a city that is nearby but in a different state. We also find that Wave-Like smoothing performs slightly better than State-level smoothing and significantly better than the Model-based smoothing due to its incorporation of multiple peaks per term, leading to more refined estimates (compared to the single peak model).

As a negative result, we can see the poor performance of model-based smoothing, which nearly undoes the positive impact of local word filtering altogether. This indicates that the model-based approach overly smooths out local perturbations in the actual data distribution, which can be useful for leveraging small local variations to locate users.

To further examine the differences among the several tested approaches, we show in Figure 2.6 the error distance in miles versus the fraction of users for whom the estimator can place within a particular error distance. The figure compares the original baseline user location estimation approach (*Baseline*), the baseline approach plus local word filtering (*+ Local Filtering*), Wave-Like smoothing approach (*LF+Wave-Like*), and then the best performing smoothing approach (*LF+Neighborhood*) and the worst performing smoothing approach (*LF+Model-based*). The x-axis identifies the error distance in miles in log-scale and the y-axis quantifies the fraction of users located within a specific error distance. We can clearly see the strong impact of local word filtering and the minor improvement of smoothing across all error dis-

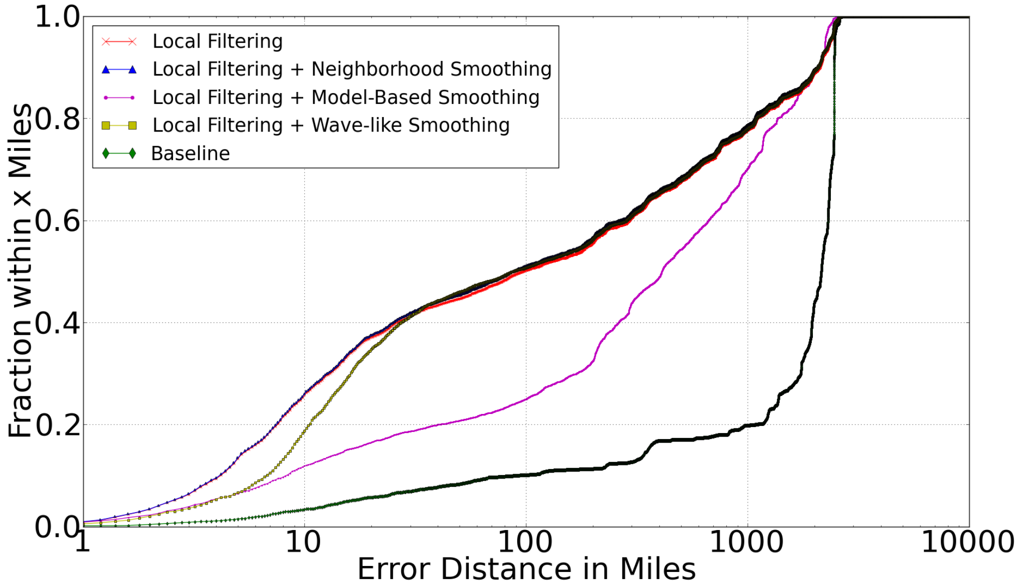


Figure 2.6: Comparison Across Estimators

tances. Interestingly, we see that the wave-like approach suffers from the problems of the model-based approach for small errors, but performs nearly as well as the neighborhood-based approach for larger errors. This suggests that the wave-like model has good potential to be further refined to eliminate the errors at small distance (introduced most likely due to the oversimplification of the model as compared to the more data-driven neighborhood-based approach). For the best performing approach, we can see that nearly 30% of users are placed within 10 miles of their actual location in addition to the 51% within 100 miles.

2.5.2 Capacity of the Estimator

To better understand the capacity of the location estimator to identify the correct user location, we next relax our requirement that the estimator make only a single location prediction. Instead, we are interested to see if the estimator can identify a good location somewhere in the top-k of its predicted cities. Such a relaxation allows us to appreciate if the estimator is identifying some local signals in many cases, even if the estimator does not place the best location in the top most probable position.

Returning to Table 2.4, we report the Accuracy@k for each of the approaches. Recall Accuracy@k measures the fraction of users located within 100 miles of their actual location, for some city in the top k predictions of the estimator. For example, for Accuracy@5 for *LF+Neighborhood* we find a result of 0.788, meaning that within the first five estimated locations, we find at least one location within 100 miles of the actual location in 79% of cases. This indicates that the content-based location estimator has high capacity for accurate location estimation, considering the top-5 cities are recommended from a pool of all cities in the US. This is a positive sign for making further refinements and ultimately to improving the top-1 city prediction.

Similarly, Figure 2.7a shows the error distance distribution for varying choices of k, where each point represents the fraction of users with an error in that range (i.e., the first point represents errors of 0-100 miles, the second point 100-200 miles, and so on). The location estimator identifies a city in the top-10 that lies within 100 miles of a user's actual city in 90% of all cases. Considering the top-1, top-3, top-5, and top-10, we can see that the location estimator performs increasingly well. Figure 2.7b continues this analysis by reporting the Average Error Distance as we consider increasing k. The original reported error of around 500 miles for the top-1 prediction drops as we increase k, down to just 82 miles when we consider the best possible city in the top-10.

2.5.3 Estimation Quality: Number of Tweets

An important question remains: how does the quality of estimation change with an increasing amount of user information? In all of our experiments so far, we have considered the test set in which each user has 1000+ tweets. But perhaps we can find equally good estimation results using only 10 or 100 tweets?

To illustrate the impact of an increasing amount of user data, we begin with a specific example of a test user with a location in Salt Lake City. Figure 2.8 illustrates the sequence of city estimations based on an increasing amount of user tweet data. With 10 tweets, Chicago has the dominant highest estimated probability. With 100 tweets, several cities in California, Salt Lake City and Milwaukee exceed Chicago. By 300 tweets, the algorithm geo-locates the user in the actual city, Salt Lake City; however there is still significant noise, with several other cities ranking close behind Salt Lake City. By 500 tweets, the probability of Salt Lake City increases dramatically, converging on Salt Lake City as the user data increases to 700 tweets and then 1000 tweets.

To quantify the impact of an increasing amount of user information, we calculate the distribution of Error Distance and the Average Error Distance across all of the test users based on the Local Word filtering location estimator relying on a range of tweets from 100 to 1000. Figure 2.9a shows the error distance distribution, where each point represents the fraction of users with an error in that range (i.e., the first point represents errors of 0-100 miles, the second point 100-200 miles, and so on). The errors are distributed similarly; even with only 100 tweets, more than 40% of users are located within 100 miles. In Figure 2.9b, we can see that with only 100 tweets that the Average Error Distance is 670 miles. As more tweets are used to refine the estimation, the error drops significantly. This suggests that as users continue to tweet, they "leak" more location information which can result in more refined estimation.

2.5.4 Impact of Social Refinement

In this section, we explore the opportunity of incorporating social tie information into the content-driven location predictor (as described in Section 2.4.5). We are interested to understand whether social refinement can improve location estimation.

Using the test set described in Section 2.3, we randomly select a set of 354 users with 10 to 20 strong connected friends, where we define a strong connected friend of a user as one who is both following and followed by the user. For each of the 354 users, we crawl the user's strong connected friends, and their latest 500 tweets. In total, we have 3,137,233 tweets from 6,502 users who are strong connected friends of the 354 users. Recall that for each of the 354 users, we have the user's location in the form of latitude/longitude coordinates. Over this set of 354 users and their latest 1,000 tweets, we apply the best content-driven approach identified in the previous experiments – Local Words Filtering + Neighborhood Smoothing. We find that this content-driven approach (with no social refinement) results in an accuracy of 54.26% and an average error distance of 472.26 miles.

2.5.4.1 Quality of Estimator: Varying α

In the social refinement algorithm, the parameter α indicates the percentage of location estimate information for a target user based on the social ties of the target user versus the target user's own content. A value of 1.0 for α means the prediction is totally based on a user u's social relations, without any input from the user u's own tweets. On the other hand, a value of 0.0 for α means the prediction is based only on user u's tweets. We tune the parameter α from 0.0 to 1.0 with an interval of 0.1 to study to what extent more information from a user's social relations can help locate the target user. In Figure 2.10a, the result shows that although we can get the highest accuracy either with an α value of 0.0 or 0.2, generally higher weights of social refinement (i.e., larger α values) produce worse results in terms of accuracy. Similarly, we show results for average error distance over different values for parameter α in Figure 2.10b. The best average error distance we get is 466.20 miles with α value 0.2, which is a 1.28% increase over the non-social refinement based algorithm (472.26 miles). Interestingly, we see the same trend that incorporating some additional evidence from a target user's social ties results in better location estimation, but over-reliance on social ties results in poorer location estimation. Surprisingly, even in the extreme case when none of a target user's content is used for location estimation (when $\alpha = 1.0$), the social ties alone still yield an estimate that is within 10% of the case when the target user's content is actually included in the estimator.

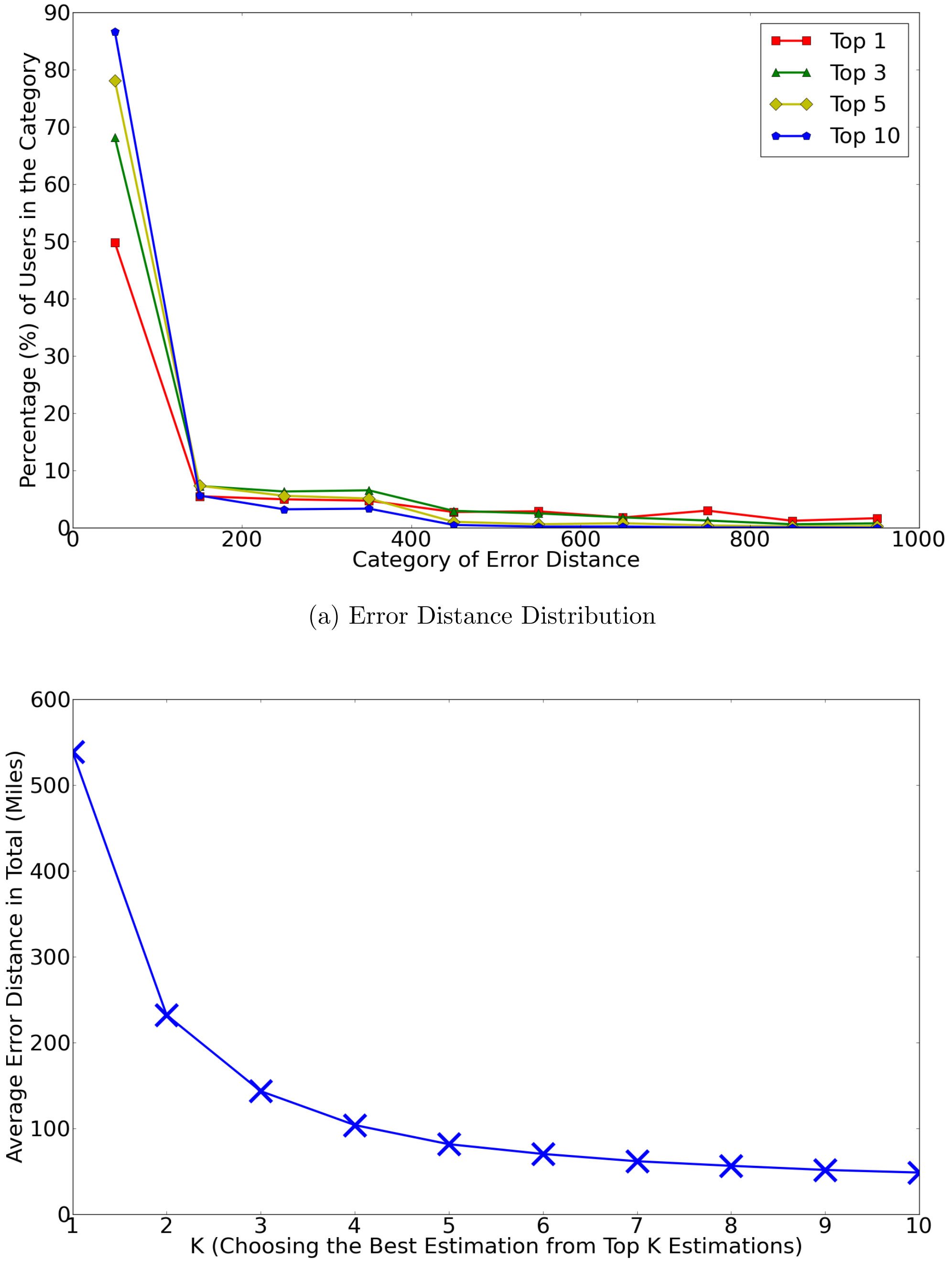
2.5.4.2 Quality of Estimator: Number of Tweets

In the second study of social refinement, we consider the impact of knowing more about the target user via additional tweets. Fixing $\alpha = 0.2$ based on the results from the previous experiment, we fix the number of tweets per social tie at 500, but vary the number of tweets for the target user from 0 to 1,000. Again, we see that even when no content is available for the target user, that the social-based estimator still achieves reasonable results (46% of users with error distance less than 100 miles; average error distance of 466 miles). As the amount of content for the target user increases, Figure 2.11a shows the improvement in accuracy, ultimately achieving around 54% accuracy. Similarly, Figure 2.11b shows how – after an initial increase from using a target user's own content – that additional content from the target user results in an improved average error distance, which echoes the results described in our location estimation experiments without social refinement (recall Figure 2.7).

Together, these experiments on social refinement of content-driven location estimation suggest a great possibility for propagating estimated user locations along social ties to target user's for whom we have no (or little) content information. We are also interested to explore more sophisticated variations of the social refinement algorithm, for example, by selectively considering only neighbors of a target user for whom we have high confidence (rather than including all neighbors) as well as PageRank-style iterative refinement approaches that aggregate not just one-hop social ties, but consider multi-hop social ties.

2.6 Summary

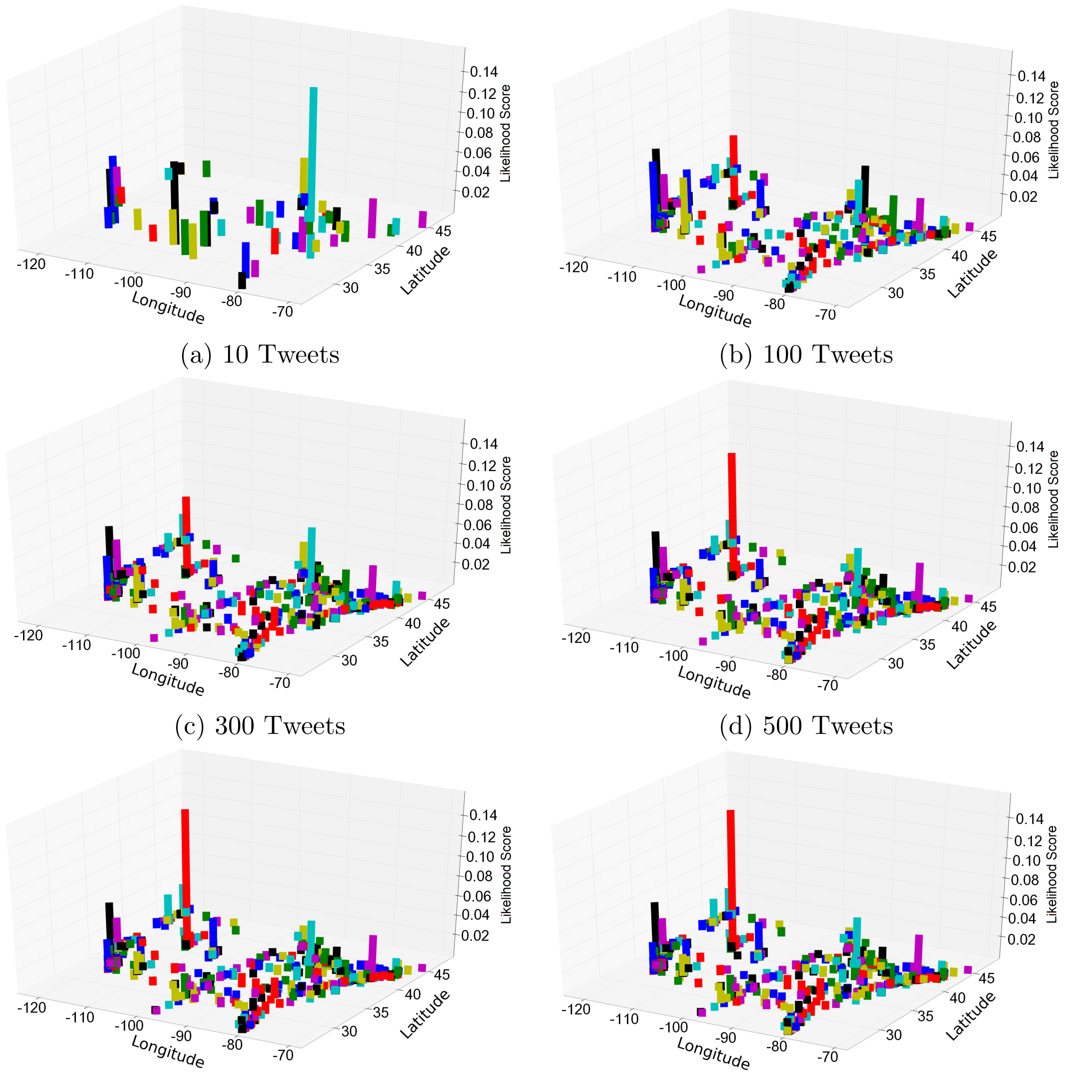
The promise of the massive human-powered sensing capabilities of Twitter and related microblogging services depends heavily on the presence of location information, which we have seen is largely absent from the majority of Twitter users. To overcome this location sparsity and to enable new location-based personalized information services, we have proposed and evaluated a probabilistic framework for estimating a microblog user's city-level location based purely on the publicly available content of the user's posts, even in the absence of any other geospatial cues. The content-based approach relies on two key refinements: (i) a classification component for automatically identifying words in tweets with a strong local geo-scope; and (ii) a lattice-based neighborhood smoothing model for refining a user's location estimate. We have seen how the location estimator can place 51% of Twitter users within 100 miles of their actual location.



(a) Error Distance Distribution

(b) Average Error Distance

Figure 2.7: Capacity of the Location Estimator: Using the Best Estimation in the Top-k



(a) 10 Tweets

(b) 100 Tweets

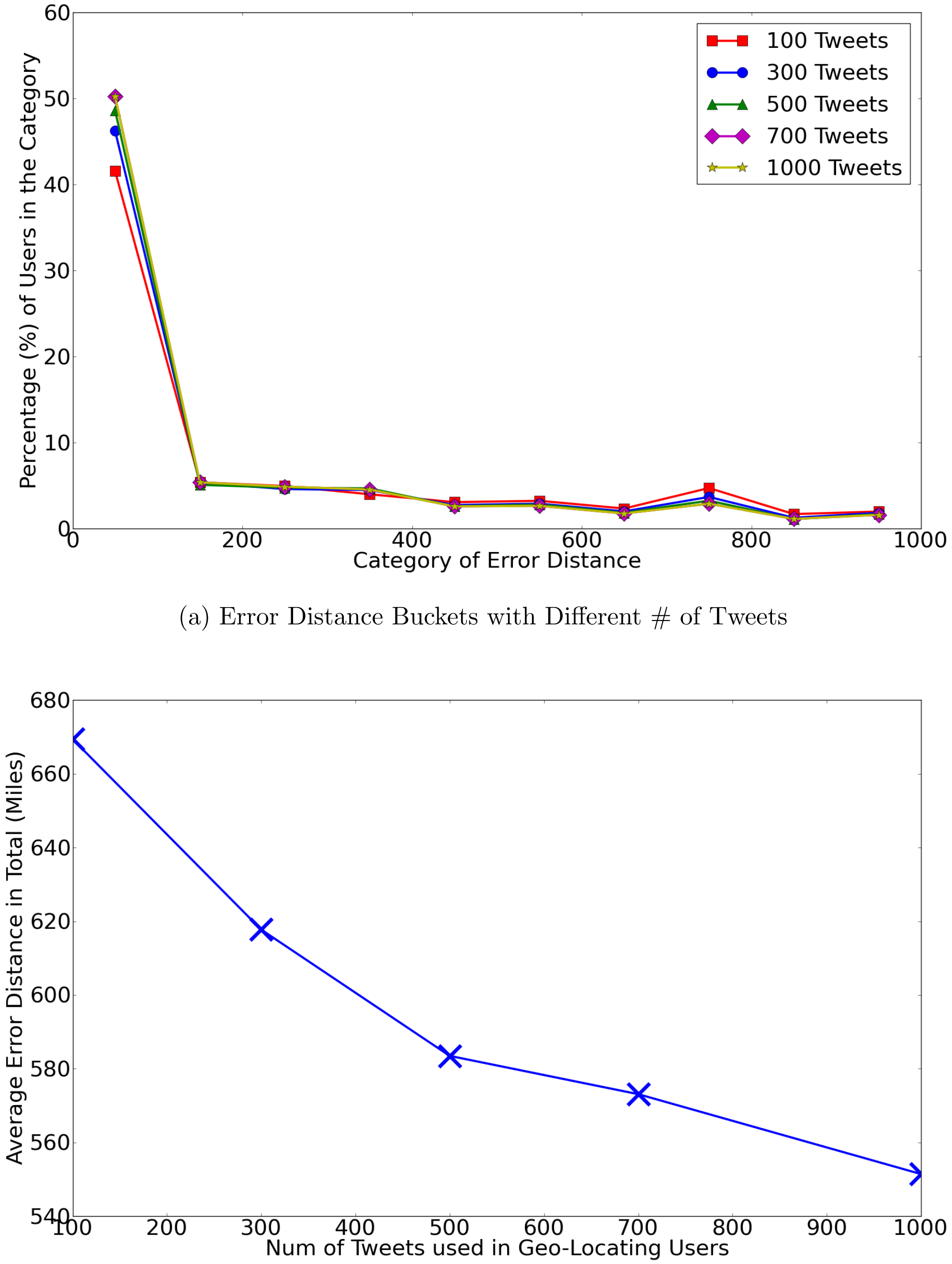
(c) 300 Tweets

(d) 500 Tweets

(e) 700 Tweets

(f) 1,000 Tweets

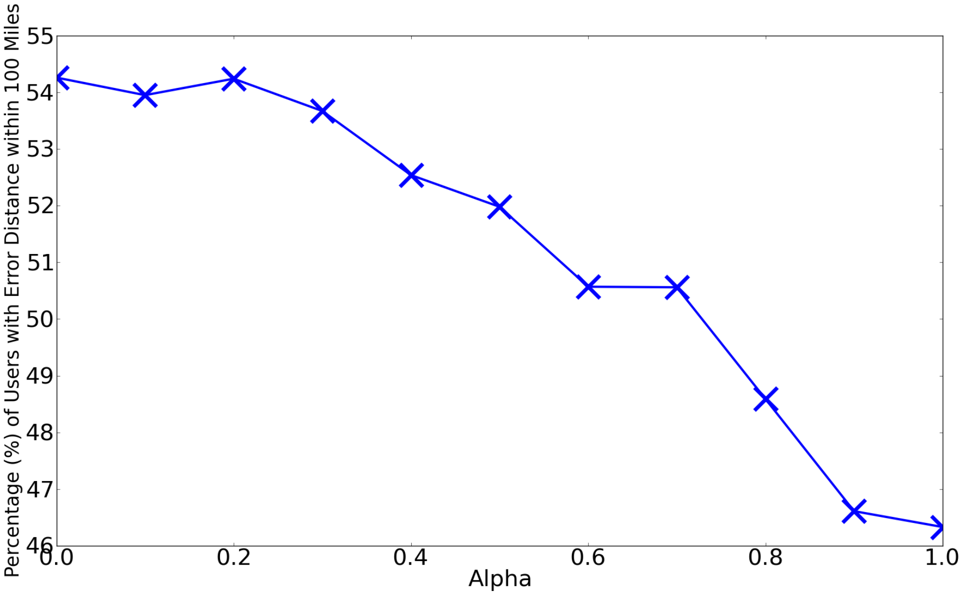
Figure 2.8: Example: Location Estimation Convergence as Number of Tweets Increases



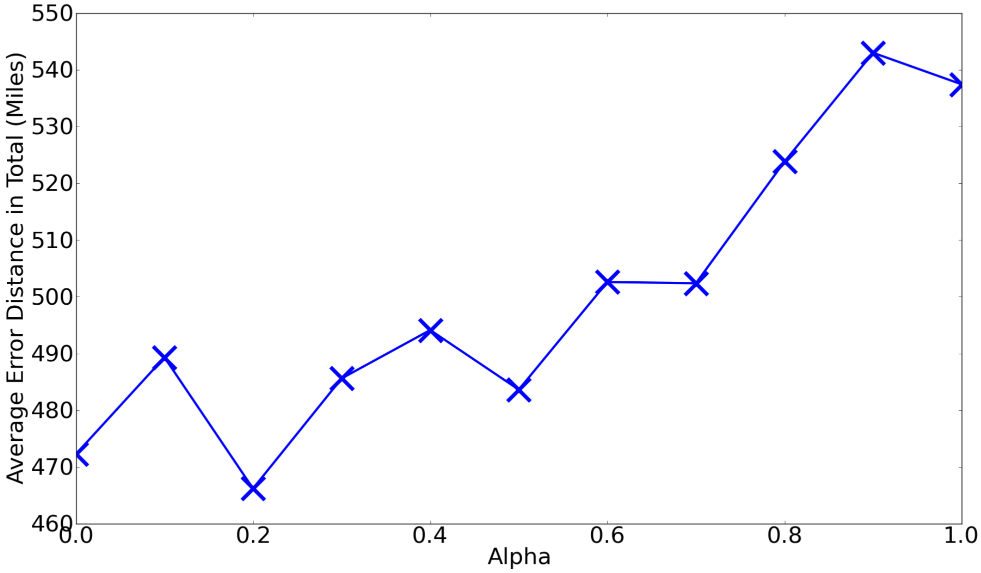
(a) Error Distance Buckets with Different # of Tweets

(b) Average Error Distance with Different # of Tweets

Figure 2.9: Refinement of Location Estimation with Increasing Number of Tweets

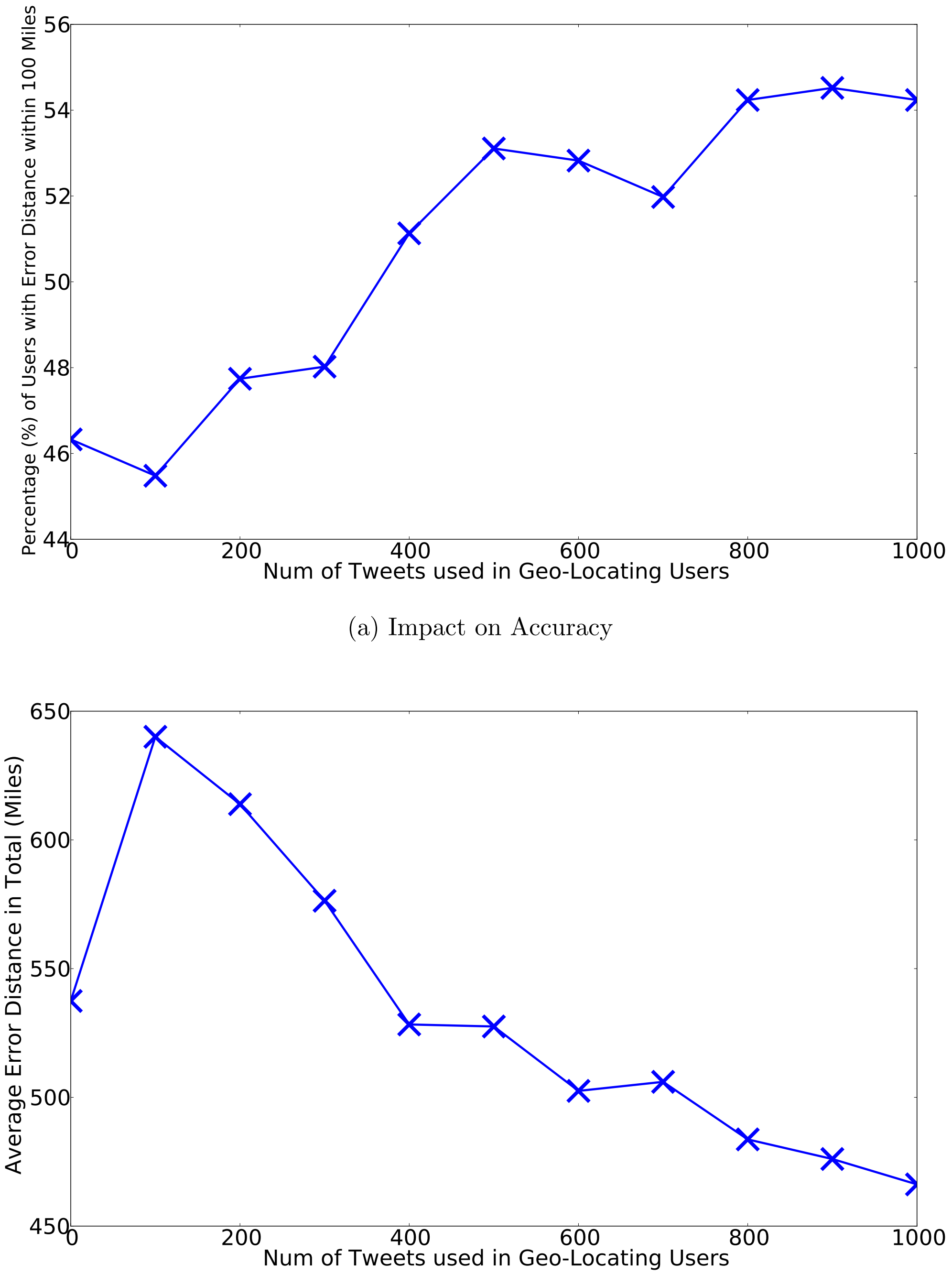


(a) Impact on Accuracy



(b) Impact on Average Error Distance

Figure 2.10: Quality of Socially Refined Estimator: Tuning Parameter α



(a) Impact on Accuracy

(b) Impact on Average Error Distance

Figure 2.11: Capacity of the Socially Refined Estimator: Varying the Number of Tweets from the Target User

3. WHOS, WHATS, AND WHENS OF LOCATION SHARING[*]

In this section, we tackle the challenge of lack of understanding of the properties of people's geo-social footprints from location sharing services. Specifically, we investigate 22 million check-ins across 220,000 users and report a quantitative assessment of human mobility patterns by analyzing the spatial, temporal, social, and textual aspects associated with these footprints.

3.1 Introduction

We are beginning to see a similar rise of location sharing services like Foursquare, Facebook Places, and Google Local. As the one of the front runners for location sharing services, the Foursquare service alone claims over 40 million registered users, and over 4.5 billion check-ins in total, with millions more every day [40]. Since Twitter started to support location geotagging associated with tweets since 2009, Twitter boasts a collection of billions of geotagged tweets [97].

Like similar services, Foursquare allows users to "check in" at different venues (e.g., grocery stores, restaurants), write tips, and upload pictures and videos. While users of Foursquare and related location sharing services may not be a representative cross-section of the whole human society, the data revealed through these services provides a fascinating and unique opportunity to study large-scale voluntarily contributed human mobility data, which could impact the design of future mobile+location-based services, traffic forecasting, urban planning, and models of disease spread.

Toward understanding the spatial, temporal, and social characteristics of how

[*] Reprinted with permission from "Exploring Millions of Footprints in Location Sharing Services" by Zhiyuan Cheng, James Caverlee, Kyumin Lee, and Daniel Z. Sui, 2011. *in Proceedings of the 5th International AAAI Conference on Weblogs and Social Media.* Copyright 2011 by AAAI.

people use these services, we present in this section a large-scale study of location sharing services. Concretely, we study the wheres and whens of over 22 million check-ins across the globe. We study human mobility patterns revealed by these check-ins and explore factors that influence this mobility, including social status, sentiment, and geographic constraints.

3.2 Related Work

The role of geography and location in online social networks has recently attracted increasing attention. Facebook researchers analyzed the distance between Facebook users' social relations, and utilized locations of a user's friends' to predict the user's geographical location [7]. Characterizing network properties in relation to local geography is studied in [126]. User behavior with regard to the location field in Twitter user profiles has been studied in [53]. [75] analyzed how and why people use location sharing services, and discussed the privacy issues related to location sharing services. Besides locations, researchers have also explored temporal dynamics associated with on-line social activities [47].

Analyzing and modeling mobility patterns has long attracted attention by researchers in fields like statistical physics, ubiquitous computing, and spatial data mining. For example, an analysis of 100,000 cellphone users' trajectories [48] showed that human mobility displayed simple reproducible patterns. The authors of [13] analyzed the circulation of bank notes in the US and concluded that human traveling behavior can be described mathematically on many spatio-temporal scales by a two parameter continuous time random walk model. A 93% potential predictability in user mobility was found across 50,000 cellphone users in [109]. [136] proposed a system to mine interesting locations and travel sequences from users' GPS trajectories. Researchers of [55] observed Lèvy Flight search patterns across 14 species of ma-

Table 3.1: Distribution of Sources of Check-ins

Name	Percentage
Foursquare	53.5%
UberTwitter	16.4%
Twitter for iPhone	10.2%
Twitter for Android	3.4%
TweetDeck	3.1%
Gowalla	2.9%
Echofon	2.0%
Gravity	1.3%
TwitBird	1.1%
Others	6.0%

rine predators, with a few individuals switching between Lèvy Flight and Brownian motion as they traversed different habitat types.

Different from cellphone data and trajectories derived from GPS trackers, check-ins have several unique features: (i) they are inherently social, since users reveal their location to their friends, meaning that social structure and its impact on human mobility can be directly observed; (ii) check-ins are associated with particular venues (e.g., a restaurant), allowing for greater analysis of venue type; (iii) check-ins can be augmented with short messages, providing partial insight into the thoughts and motivations of users of these services.

3.3 Gathering Check-ins

To begin our study, we first require a collection of check-ins. Since personal check-in information on location sharing services like Foursquare, Gowalla, and Facebook Places is typically restricted to a user's immediate social circle (and hence unavailable for sampling) we take an approach in which we sample location sharing status updates from the public Twitter feed. Twitter status messages support the inclusion of geo-tags (latitude/longitude) as well as support third-party location sharing services like

Foursquare and Gowalla (where users of these services opt-in to share their check-ins on Twitter). We monitor Twitter's gardenhose streaming API (∼1% of the entire Twitter public timeline), and retrieve users who post geo-tagged status updates. For each sampled user, we crawl up to a maximum of the most recent 2,000 geo-labeled tweets.

The location crawler ran from late September 2010 to late January 2011, resulting in a total collection of 225,098 users and 22,506,721 unique check-ins. The 22 million check-ins were posted from more than 1,200 applications, and the distribution of sources is displayed in Table 3.1. More than 53% of the check-ins are from Foursquare, and most of the other check-ins are from Twitter's applications on mobile platforms like Blackberry, Android, and iPhone. A few hundred thousands check-ins are from other location sharing services like Gowalla, Echofon, and Gravity.

3.3.1 Format of the Data

Each check-in is stored as the tuple *checkin(userID, tweetID)* = {*userID, tweetID, text, location, time, venueID*}. An example check-in tuple is: checkin(14091113, 9710376274) = {14091113, 9710376274, "I'm at MTA - Atlantic Ave-Pacific St Subway Station. http://4sq.com/2nWVD0", 40.685307, -73.980719, "2010-02-26 21:42:04", "cd979d2e352c4f54"}. We additionally store a user as the tuple: *user(userID)* = {*userID, status_count, followers_count, followings_count*}; for the example check-in, the user has 2,771 total status updates, 255 followers and is following 926 users.

3.3.2 Filtering Noise

Many location sharing services provide some mechanism to verify that a user is actually at or near the venue where they are checking in (e.g., by cross-checking with a user's cellphone GPS) [39], however, there can still be incidents of false check-ins. Hence, we additionally filter out all check-ins from users whose consecutive check-ins

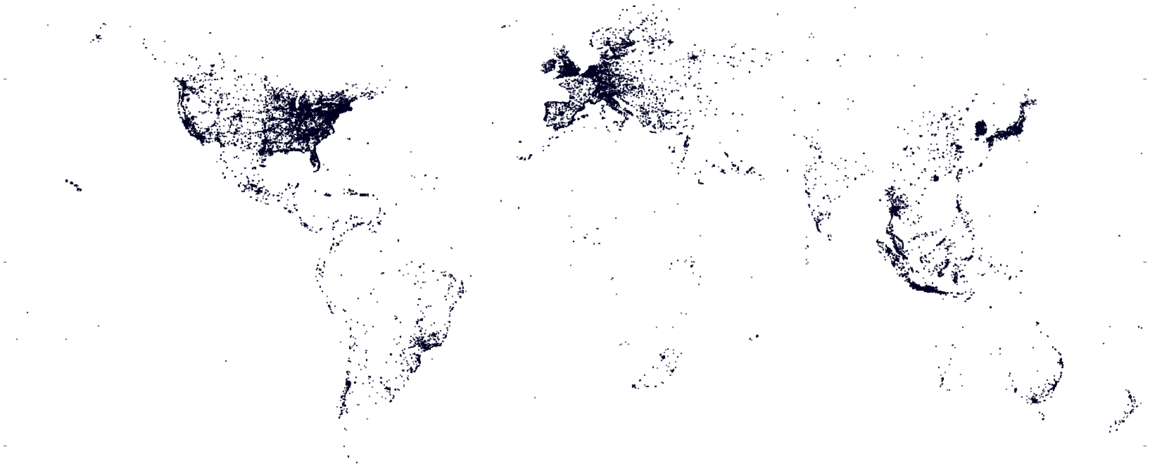


Figure 3.1: Global Distribution of Check-ins

imply a rate of speed faster than 1000 miles-per-hour (or faster than an airplane). In total, we filtered 294 users (0.1%) with sudden moves, yielding a final collection of 224,804 users and 22,388,315 check-ins. More than 72% users have fewer than 100 check-ins; 7.8% users have more than 300 check-ins; and 3.6% users have more than 500.[†]

3.3.3 Locating Each User's "Home"

Some of the analysis in the following sections requires that we first associate each user with a natural "home", so, for example, we can compare the properties of all users "from" New York City versus users "from" Los Angeles. Since users of location sharing services are not required to register a home location, we must algorithmically determine the home location. Note that choosing a user's home based on the center of mass of all check-ins suffers from splitting-the-difference, by placing a user from Houston who occasionally travels to Dallas somewhere in between the two cities; alternatively, directly considering the user's most frequently checked-in venue may

[†] Data are available at http://infolab.tamu.edu/data/

overlook a cluster of closely-located but less individually checked-in venues. To avoid these drawbacks, we propose a simple method to geo-locate a user's home based on a recursive grid search. First, we group check-ins into squares of one degree latitude by one degree longitude (covering about 4,000 square miles). Next, we select the square containing the most check-ins as the center, and select the eight neighboring squares to form a lattice. We divide the lattice into squares measuring 0.1 by 0.1 square degrees, and repeat the center and neighbor selection procedures. This process repeats until we arrive at squares of size 0.001 by 0.001 square degrees (covering about 0.004 square miles). Finally, we select the center of the square with the most check-ins as the "home" of the user.

3.4 Spatio-Temporal Analysis of Check-ins

In this section, we begin our study of large-scale location sharing services with an investigation of the temporal and geographic characteristics of how people use these services.

3.4.1 Wheres of the Check-ins

First, we plot the locations of the 22 million check-ins in Figure 3.1, where we see that while check-ins are globally distributed, the density of check-ins is highest in North America, Western Europe, South Asia, and Pacific Asia. Zooming in on the US, Figure 3.2 shows the reach of location sharing services, revealing the boundaries of cities and the lines of highways. Further zooming in, we can see in Figure 3.3 how New York City is densely covered by more than half a million check-ins. While these figures convey the scale and density of location sharing services, we can further explore the nature of these check-ins by aggregating keywords across all 22 million check-in tuples. The aggregated view in Figure 3.4 shows that the most popular check-in venues are restaurants, coffee shops, stores, airports, and other venues re-

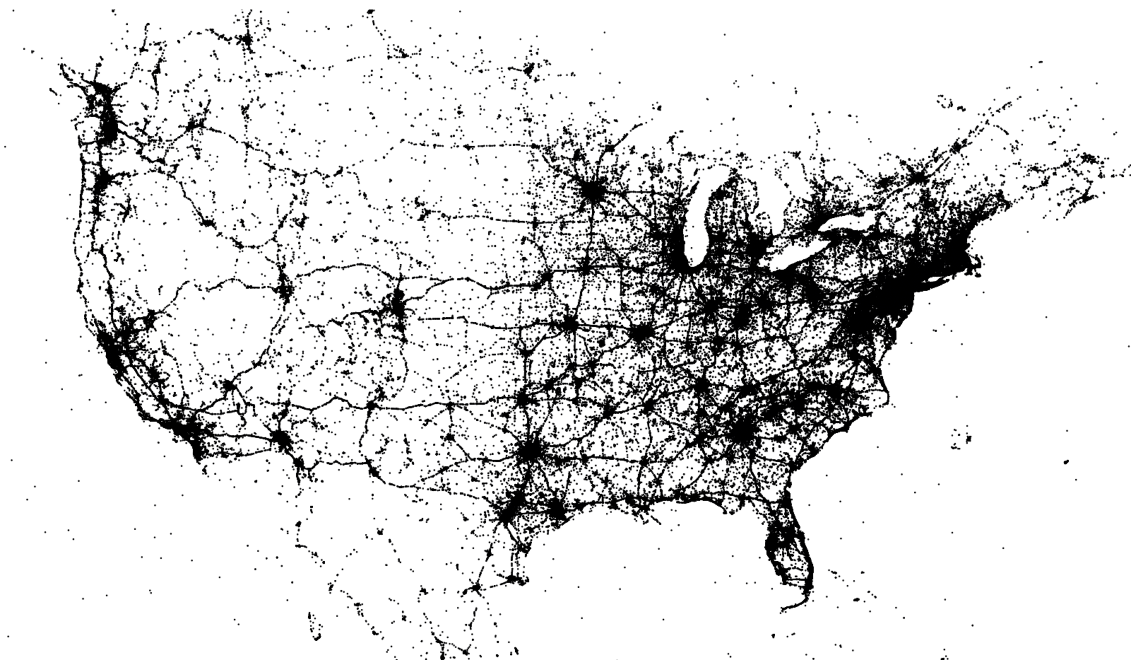


Figure 3.2: Detail: Check-ins in the United States

flecting daily activity (e.g., fitness, pubs, church).

3.4.2 Whens of the Check-ins

Considering the temporal distribution of check-ins, we can uncover both the aggregate daily patterns of users of location sharing services and their weekly patterns. By normalizing the timestamps of every check-in so that all local times are treated as the same time (i.e., aggregating all check-ins at 1pm, whether they be in Chicago or Tokyo), we show in Figure 3.5 the mean check-in pattern per day. This pattern provides a glimpse into the global daily "heartbeat", with three major peaks: one around 9am, one around 12pm, and one around 6pm. The diurnal pattern is clearly displayed as more people are active during the daytime than at night.

To illustrate the potential of location sharing services as sociometers of city health and activity, we show in Figure 3.6, the disaggregated daily check-in patterns of users

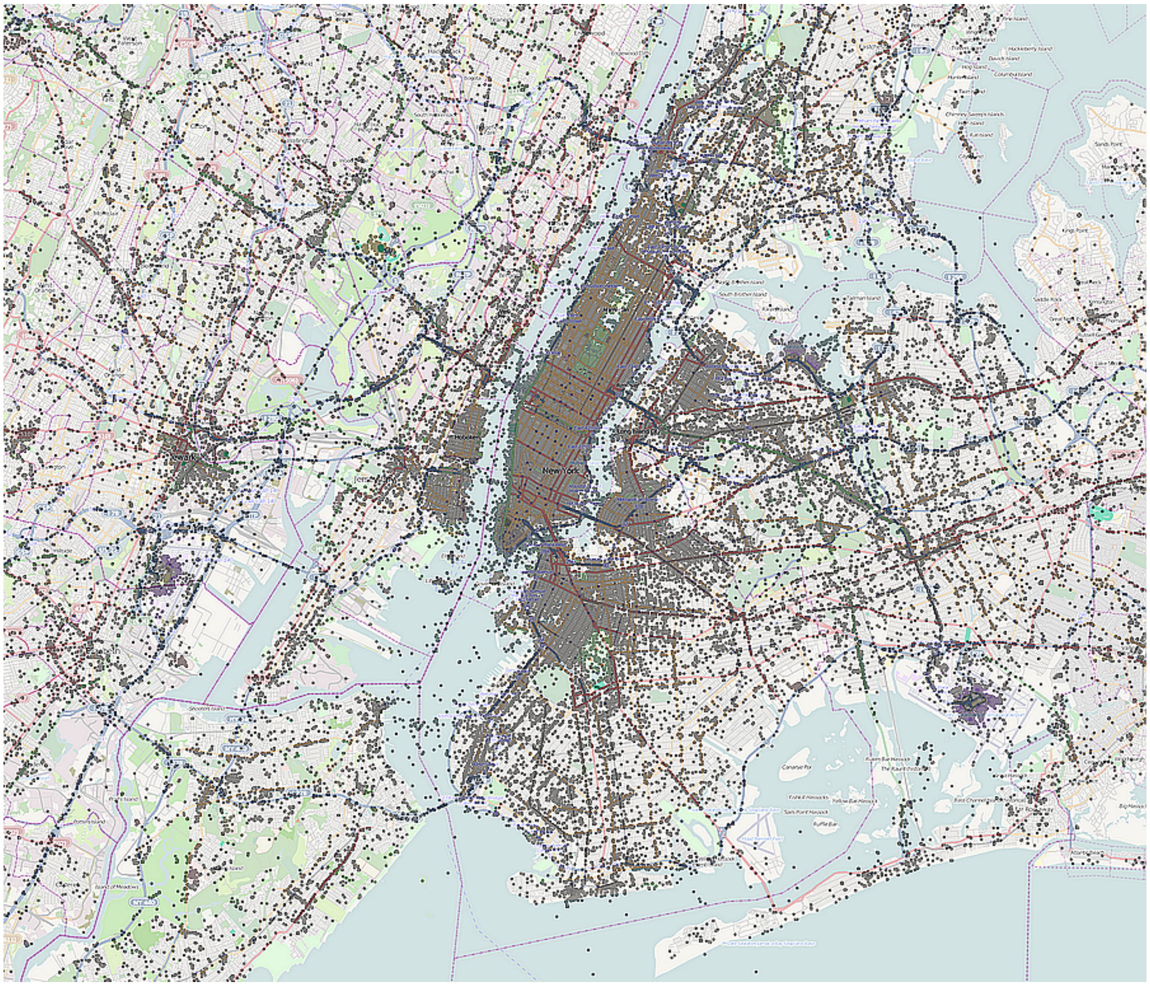


Figure 3.3: Detail: Check-ins in New York City

in New York City, Los Angeles, and Amsterdam. The check-in patterns show that Amsterdam's daily "heartbeat" reflects an early-rising city, with more activity than either LA or New York in the morning hours. LA peaks around noon, whereas New York has the highest check-in rate during the night ("The City That Never Sleeps"). We are interested to further explore the reasons for these differences. Are the daily differences artifacts of local culture? Or the proclivity of users in certain locations to more willingly reveal certain aspects of their daily lives than others (e.g., check-in in while at work, but not at play?) Or do the differences reflect biases in the data,

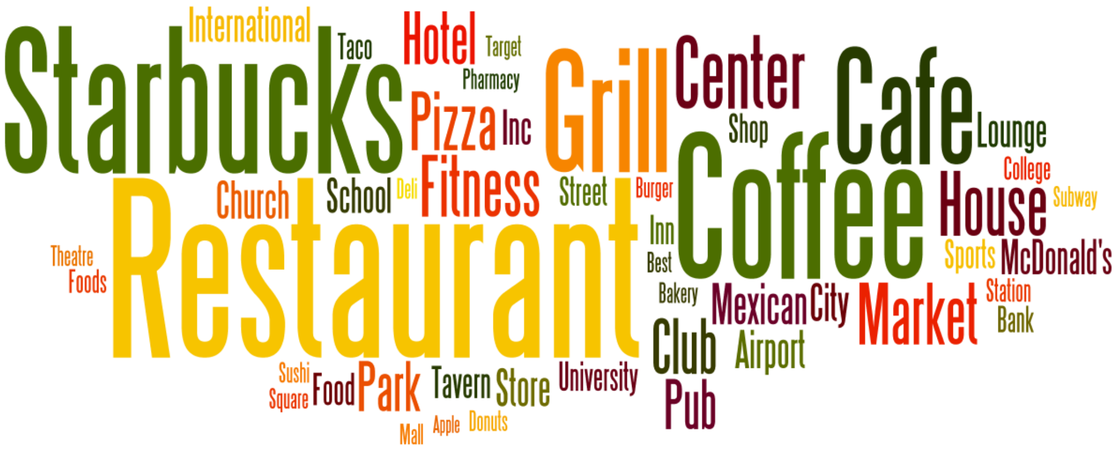


Figure 3.4: Venue Cloud for Check-ins

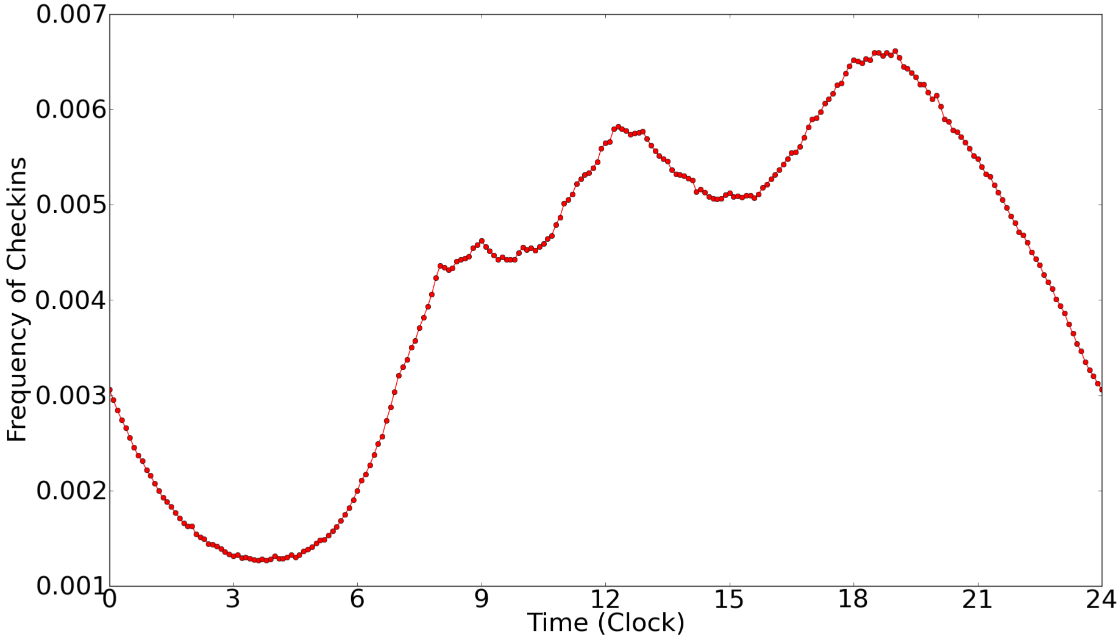


Figure 3.5: Mean Daily Checkin Pattern

so that certain demographics are over-represented in one city versus another?

Moving from the daily pattern to the weekly pattern, we see in Figure 3.7 the aggregate global patterns over the days of the week. Weekdays clearly indicate two

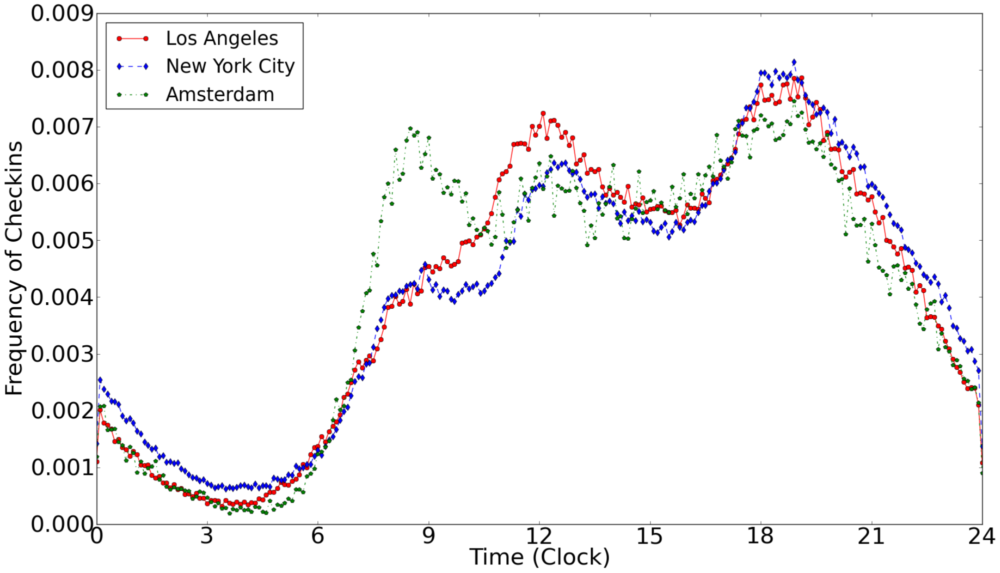


Figure 3.6: Daily Checkin Patterns: NYC, LA, Amsterdam

peaks during lunch time and dinner time, while over the weekend these two peaks blend, reflecting a fundamentally different weekend schedule for most users of location sharing services. We can also observe that the relative daily activity increases from Monday to Friday, peaking on Friday evening.

3.5 Studying Human Mobility Patterns

Given the global coverage of location sharing services and the potential of user check-ins to reveal temporal patterns of human behavior, we next turn to an examination of mobility patterns reflected in the check-in data. We consider three statistical properties often used in the study and modeling of human mobility patterns – displacement, radius of gyration, and returning probability. Taken together, these properties can inform whether humans follow simple reproducible patterns, and can have a strong impact on all phenomena driven by human mobility, from epidemic prevention to emergency response, urban planning and agent-based modeling.

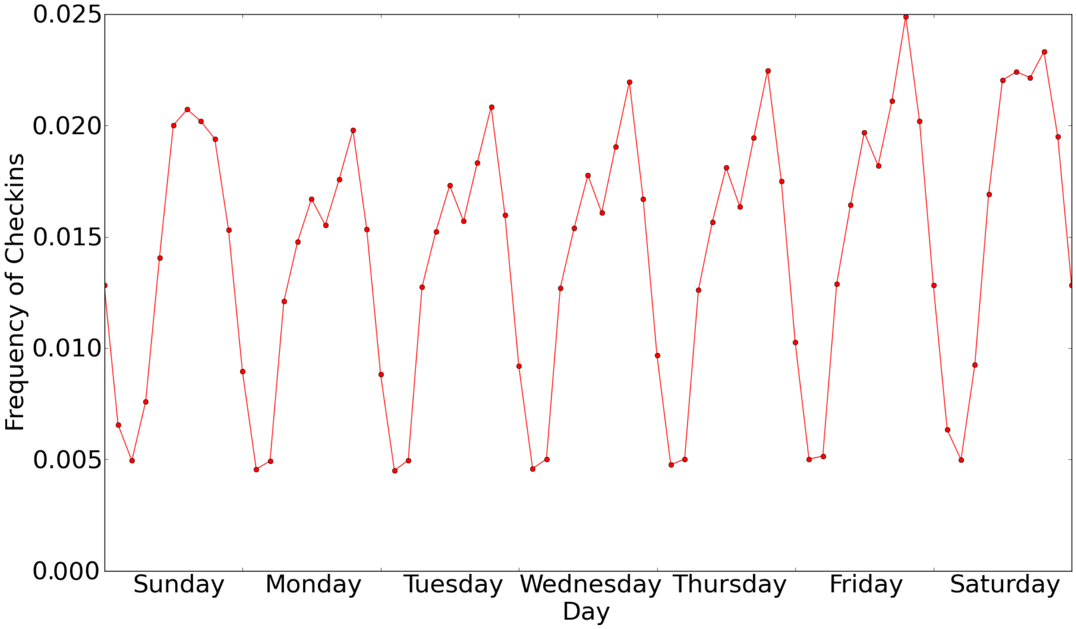


Figure 3.7: Mean Weekly Checkin Pattern

3.5.1 User Displacement

We begin with an investigation of the distance-based *displacement* of consecutive check-ins made by users. Considering all pairs of consecutive check-ins yields 22,163,511 separate displacements, reflecting the distance between these consecutive check-ins (and hence, how far a user has traveled). We plot the distribution of displacement for the dataset on a log-log scale in Figure 3.8. The x-axis is the displacement in miles, and the y-axis is the frequency of displacements in the same bucket. The trend is approximated by a power-law:

$$P(\delta_r) \propto \delta_r^{-\beta}$$

where δ_r represents the displacement and $\beta = 1.8845$. The formula indicates that human motion modeled with check-in data follows a Lévy Flight [96], in which a

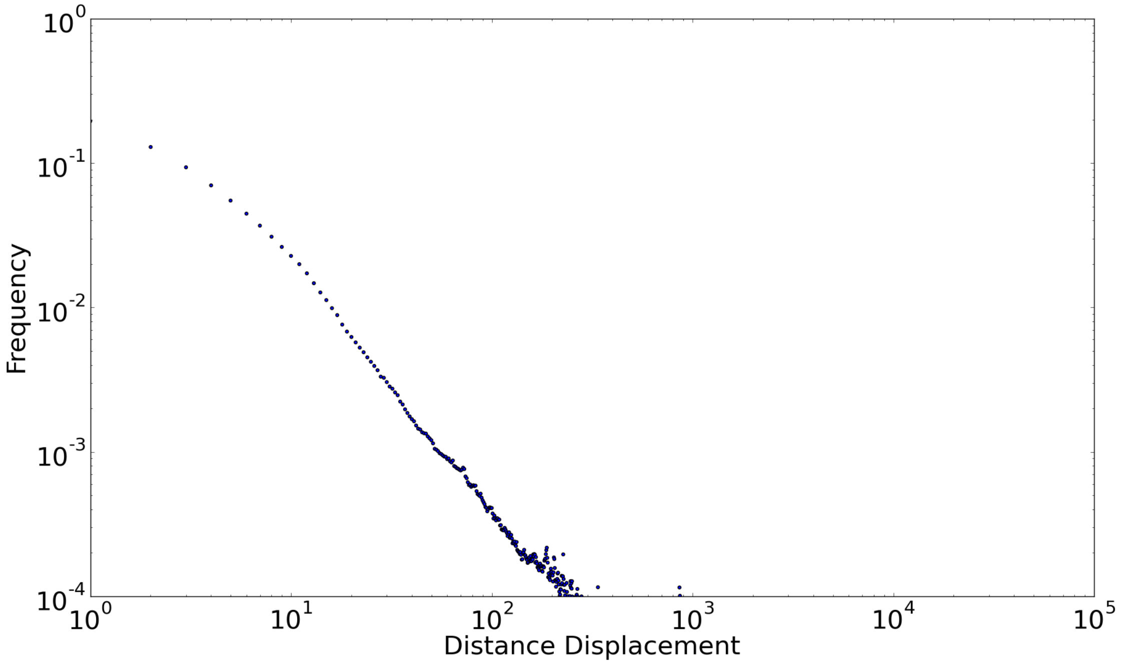


Figure 3.8: Distribution of Displacements

random walk proceeds according to steps drawn from a heavy-tailed distribution. A Lévy Flight is characterized by a mixture of short, random movements with occasional long jumps. Flight models with a similar scaling exponent have been observed separately in a study of displacements based on cellphone call data with $\beta = 1.75$ [48] and in a study of displacements based on bank note dispersal with $\beta = 1.59$ [13].

3.5.2 Radius of Gyration

Second, we consider the *radius of gyration* of each user, which measures the standard deviation of distances between the user's check-ins and the user's center of mass. The radius of gyration measures both how frequently and how far a user moves. A low radius of gyration typically indicates a user who travels mainly locally (with few long-distance check-ins), while a high radius of gyration indicates a user with many long-distance check-ins. The radius of gyration for a user can be formalized

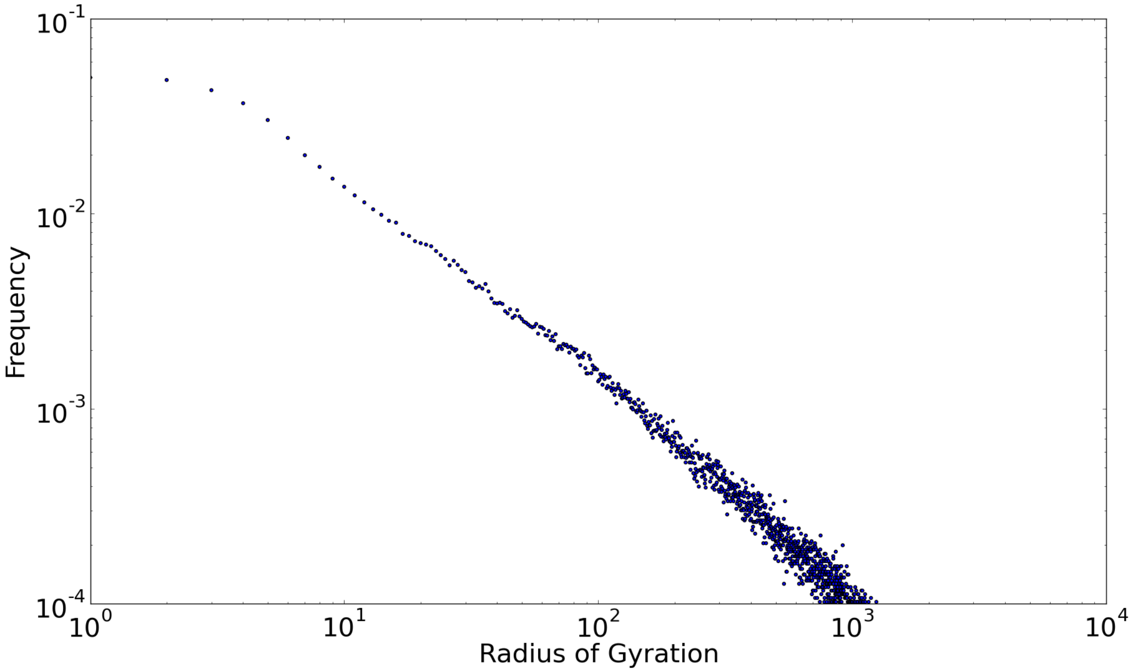


Figure 3.9: Distribution of Radius of Gyration

as:

$$r_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i - r_{cm})^2}$$

where n is the number of check-ins of the user, and $(r_i - r_{cm})$ is the distance between a particular check-in r_i and the user's center of mass r_{cm} (which is a simple average location over all check-ins). We calculate the radius of gyration for each user in our collection and the distribution of radius of gyration is displayed on A log-log scale in Figure 3.9. The x-axis identifies the radius of gyration in miles and the y-axis shows the number of users with that radius of gyration. The trend in Figure 3.9, like the distribution of displacements, also follows a power-law:

$$P(r_g) \propto r_g^{-\beta}$$

where r_g represents the radius of gyration, and $\beta = 0.9864$. 34.5% of all users display a radius of gyration of less than 10 miles, while only 14.6% have a radius of gyration larger than 500 miles.

To illustrate how radius of gyration can give further insight into the dynamics of cities, Figure 3.10 plots the average radius of gyration of users in major cities (with 100,000+ population and at least 20 users in the check-in dataset) in the continental US. The red bubbles are cities with a radius of gyration larger than 500 miles; blue ones are cities with a radius larger than 250 miles; cyan ones have a radius larger than 125 miles, and yellow ones are the rest of major cities. Users in coastal cities tend to have a higher radius of gyration than users in inland cities, and people in central states tend to have a high radius of gyration due to long distance travels to the coasts. Even so, there are some interesting regional variations worth further study, for example, the low radius of gyration for El Paso compared to the higher radius for nearby Albuquerque.

3.5.3 Returning Probability

The third property we study – returning probability – is a measure of periodic behavior in human mobility patterns. Periodic behavior is common in people's daily life (e.g., visits to work or school every weekday; visits to the grocery store on weekends) and echoes periodic behavior observed in animal migrations when animals visit the same places at the same time each year. Do users of location sharing services display a similar periodicity?

We measure periodic behavior by the *returning probability* (or, first passage time probability), which is the probability that a user returns to a location that she first visited t hours before. Grouping all returning times of all check-ins into buckets of one-hour, we plot the distribution of returning times in Figure 3.11, in which the

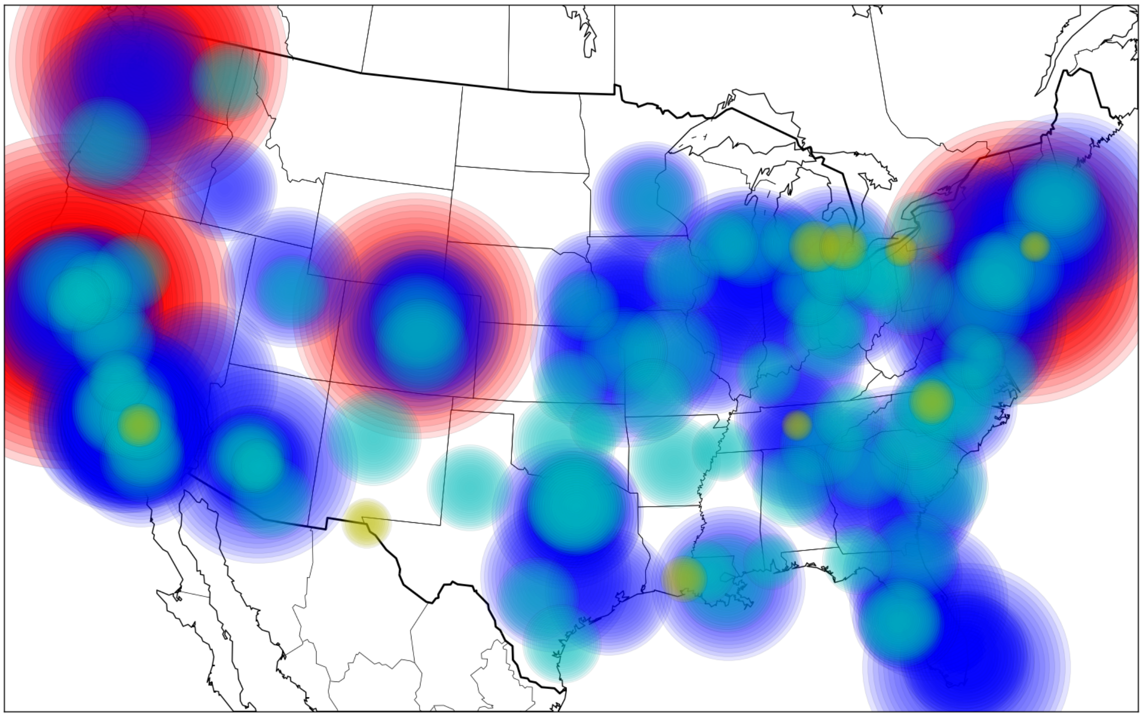


Figure 3.10: Mean Radius of Gyration for Users in US Cities

x-axis represents the bucket of returning time, and the y-axis is the corresponding frequency for a bucket. For example, at 168 hours, the returning probability peaks, indicating a strong weekly return probability. Similarly, we see daily return probabilities. As time moves forward, the returning probability shows a slight negative slope, indicating the aggregate forgetfulness of visiting previously visited places (that is, the return probability is strongest for places we have visited most recently).

3.6 Exploring Factors that Influence Mobility

In this final section, we turn our attention to exploring the factors that may impact human mobility. While factors like geography and economic status are natural to investigate, the unique properties of location sharing services provide an unprecedented opportunity to consider heretofore difficult to measure aspects of human

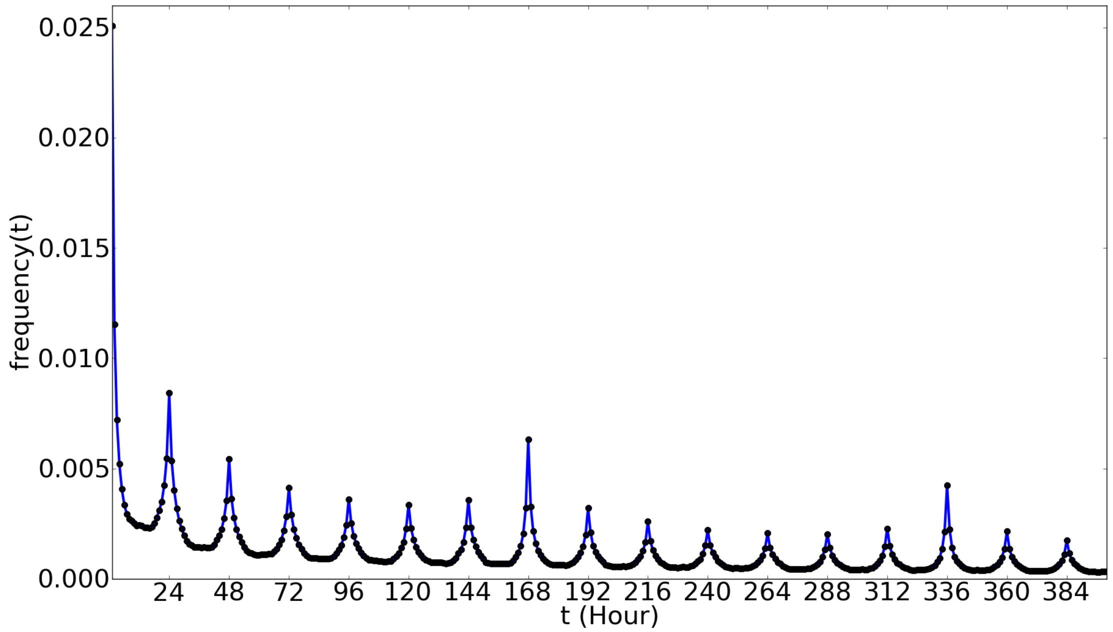


Figure 3.11: Distribution of Returning Probability

behavior. For example, does social status as measured through popularity in these services impact a user's radius of gyration? Does user-generated content implicitly reveal characteristics of the mobility of users?

3.6.1 Geographic and Economic Constraints

We begin by illustrating how geographic and economic constraints can influence human mobility patterns as revealed by location sharing services. We focus on users who are located in US cities with a population of more than 4,000. As one type of geographic constraint we consider population density and compare the radius of gyration for users from cities of differing density.[‡]

As shown in Figure 3.12, we can clearly see that people in the densest areas travel much more than people in sparse areas, but that people in the sparsest areas

[‡] Data for each US city is parsed from www.city-data.com.

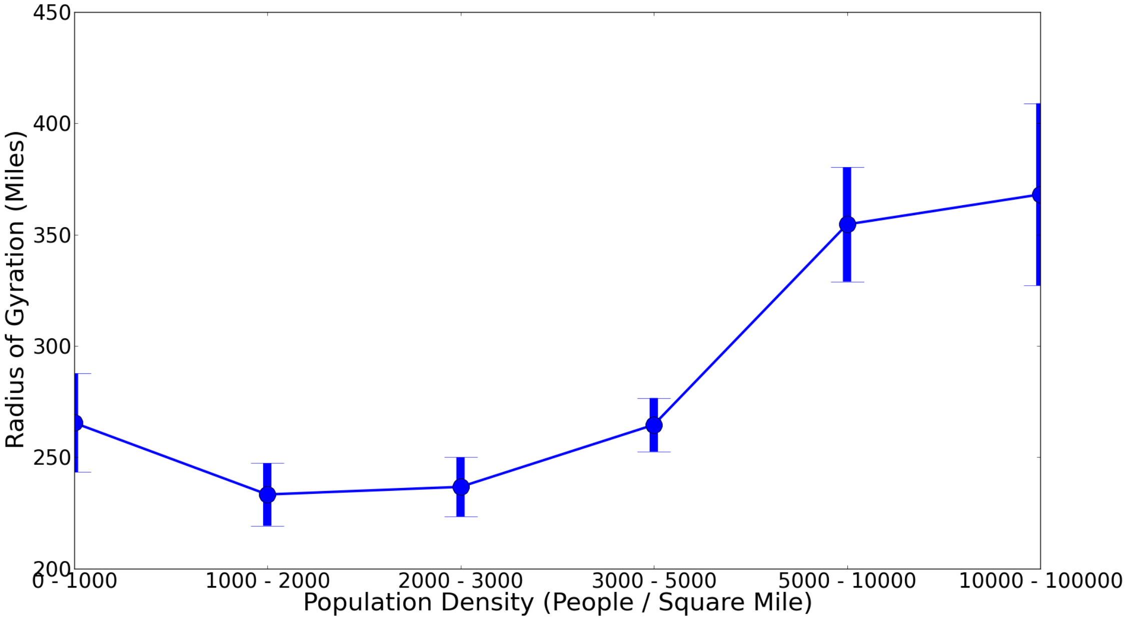


Figure 3.12: Average R_g versus City Population Density

travel farther than people in slightly denser areas. One possible explanation for both of these observations can be that: people living in metropolitan areas have more opportunities to travel for business to distant cities or countries; and people living in sparse areas (small towns) require longer travel to nearby mid-size cities.

Similarly, we can examine the economic properties of a city to understand whether economic capacity inhibits or encourages more travel by its residents. Specifically, we measure the influence of a city's average household income on its residents' radius of gyration, which is plotted in Figure 3.13. The figure shows that people in wealthy cities travel more frequently to distant places than people in less rich cities. In the meantime, people in cities with the least incomes travel slightly more than people in richer cities.

What is encouraging about both these example observations is that location sharing services provide a new window for measuring and studying fundamental proper-

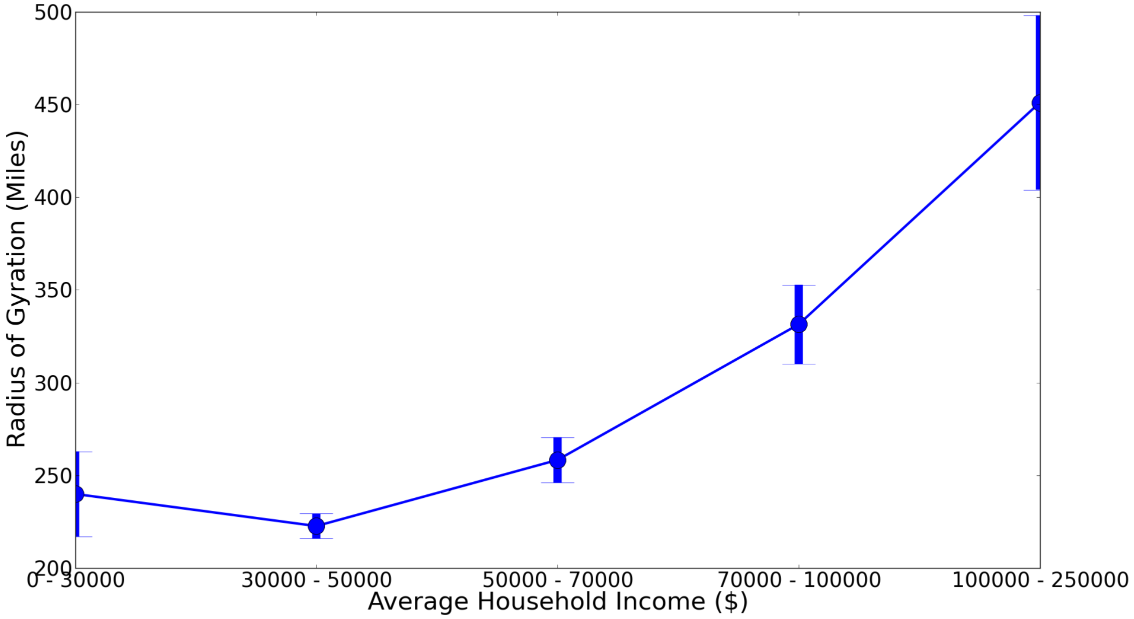


Figure 3.13: Average R_g versus City Avg Household Income

ties of cities and their residents.

3.6.2 Social Status

We next turn to one of the more exciting possibilities raised by the social structure inherent in location sharing services. Does social status impact human mobility? We consider two simple measures of status. The first is a simple measure of popularity, where we count the user's number of followers from their Twitter profile (recall the data collection method described earlier in the section; followers are one-sided friendships). The second is a measure of status that considers the ratio of a user's number of followers to the number of users that the user follows (followings):

$$status(u) = \frac{n_{followers}(u)}{n_{followings}(u)}$$

High-status users have many followers but follow very few other users themselves.

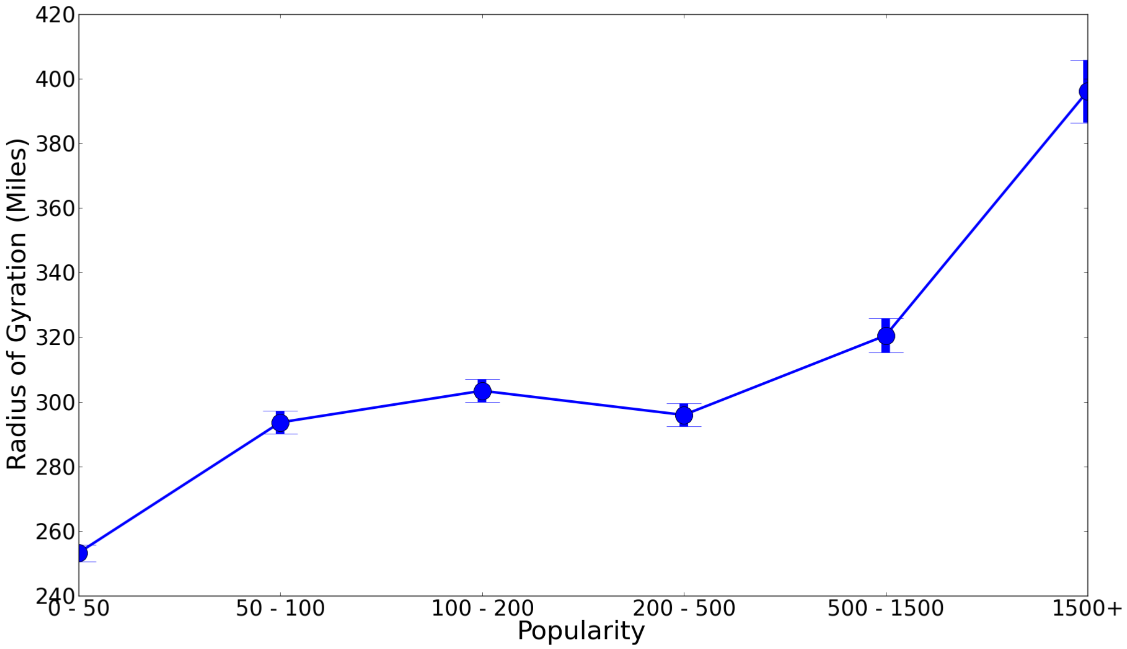


Figure 3.14: Average R_g versus Popularity

Figure 3.14 and Figure 3.15 show the relationship between both of these social status factors and the radius of gyration. We see that in both cases highly social users have higher radii of gyration than less social users. Our initial hypothesis is that users who travel have more chances to meet friends, and thus get involved in more social activities. But perhaps users with lower measured "status" engage with these social media technologies differently? For example, some Twitter users may primarily only follow other users as a form of news gathering, rather than treating Twitter as a social network of friends, resulting in lower measured status. We are interested to explore these and related questions in our ongoing research.

3.6.3 Content and Sentiment Factors

Finally, we turn to an analysis of user-generated content in location sharing services and its impact on mobility. Users of location sharing services, in addition to

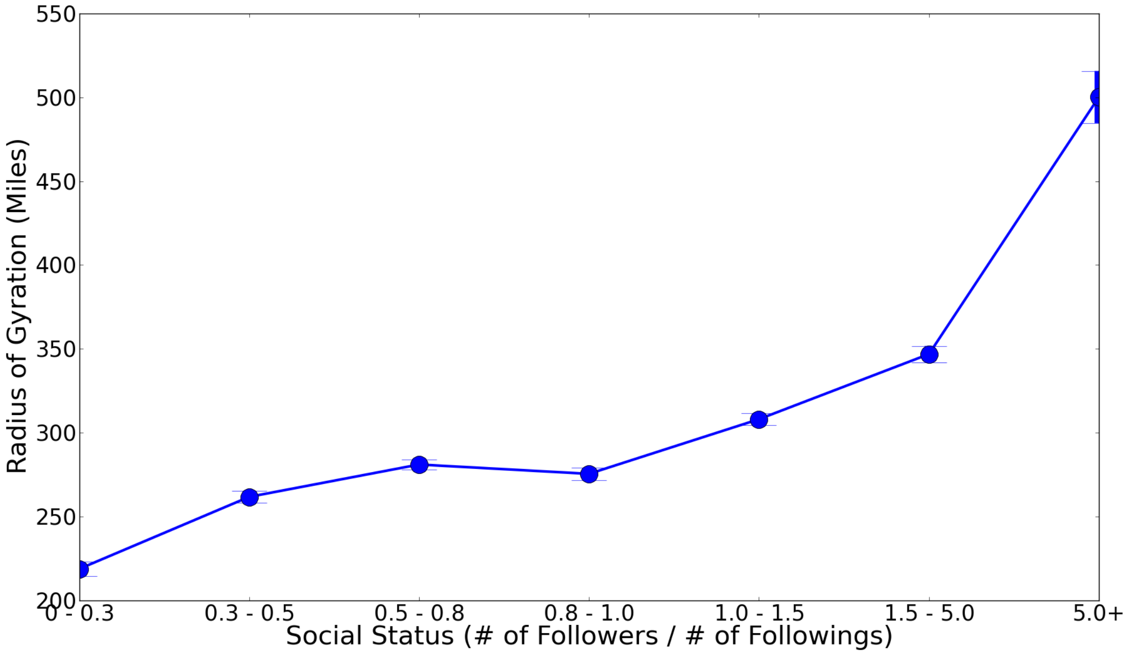


Figure 3.15: Average R_g versus Social Status

recording their location, can also post short messages, tips, and other annotations on the locations they visit. Unlike purely GPS-driven or cellphone trace data, these short messages provide a potentially rich source of context for better understanding how users engage with location sharing services.

3.6.3.1 Significant Terms vs. Radius of Gyration

Our first goal is to identify *significant terms* for users associated with varying degrees of radius of gyration, much like in our previous studies of economic, geographic, and social factors. Do high mobility users describe the world differently than low mobility users? We focus our study here on English-language messages only by using the language identification component in the NLTK toolkit [77]. We find that 49% of all users (110,559) in our collection are primarily English-language users.

To identify significant terms for these users, we identify terms with high mutual information for each category of radius of gyration. Mutual information is a standard information theoretic measure of “informativeness” and, in our case, can be used to measure the contribution of a particular term to a category of radius of gyration. Concretely, we build a unigram language model for each category of radius of gyration by aggregating all posts by all users belonging to a particular category of radius of gyration (e.g, all users with a radius of gyration between 0 and 10). Hence, mutual information is measured as: $MI(t, c) = p(t|c)p(c)log\frac{p(t|c)}{p(t)}$ where $p(t|c)$ is the probability that a user which belongs to category c has posted a message containing term t, $p(c)$ is the probability that a user belongs to category c, and $p(t)$ is the probability of term t over all categories. That is, $p(t) = count(t)/n$. Similarly, $p(t|c)$ and $p(c)$ can be simplified as $p(t|c) = count(c, t)/count(c)$ and $p(c) = count(c)/n$ respectively, where $count(c, t)$ denotes the number of users in category c which also contain term t, and $count(c)$ denotes the number of users in category c.

In Table 3.2, we report the top-10 most significant terms from users with different radii of gyration. In the table, we can clearly see the differences between frequent travelers with a large radius of gyration and the more local people with a small radius of gyration. Travelers talk a lot about long journey related terms: “international airport” (and abbreviations of international portals: “SFO”, “JFK”), major metropolitan areas (e.g., New York, San Francisco, London, Paris, Los Angeles), “flight”, and “hotel”. At lower levels of mobility, we see significant words like “railway station” and “bus”, as well as discussion of “home”, “work”, “church”, grocery stores (e.g., HEB, Walmart, “mall”), “college”, and “university”. People with different mobility patterns significantly differ in the topics they talk about and terms they use, indicating a fruitful area of further study.

Table 3.2: Top 10 Significant Terms for Each Radius of Gyration R_g Category

R_g (miles)	Top 10 Terms				
(1000,+∞)	international airport	New York	San Francisco	London	terminal
	SFO	flight	JFK	Jakarta	Paris
(500,1000]	international airport	San Francisco	New York	Las Vegas	Los Angeles
	Chicago	hotel	Seattle	terminal	Washington
(300,500]	international airport	Chicago	Dallas	New York	hotel
	Lake	Austin	Beach	Orlando	Seattle
(100,300]	airport	Chicago	Atlanta	Jakarta	hotel
	Berlin	church	center	bar	beach
(50,100]	mayor	railway station	Pittsburgh	university	Stockholm
	church	Madrid	Greenville	center	college
(10,50]	mayor	station	home	work	Bangkok
	house	HEB	school	Walmart	road
(0,10]	Singapore	home	Jakarta	Indonesia	university
	center	mall	bus	woodlands	road

3.6.3.2 Capturing User's Sentiment

We can additionally measure the relative viewpoint of users and their locations by considering the sentiment of each user's posted messages. To capture the sentiment associated with the check-ins, we use the public SentiWordNet [35] thesaurus to quantify sentiment for each English speaking user. For each message, we extract the words that have a quantified sentiment value in SentiWordNet and consider the sentiment of the post as the mean value for the sentiments for words in the post. For each user, the user's sentiment is calculated as the mean value of the sentiments of all the user's posts. In this way, we capture the sentiment for each of the 110,559 English speaking users in the dataset. The distribution of sentiment of the users is plotted in Figure 3.16, and we can clearly see that most users have a neutral sentiment, and only a small portion of users express strong sentiment when using location sharing services.

When we drill down to see which words are associated with a positive, neutral, and negative sentiment (again, using mutual information) we see in Table 3.3 that most

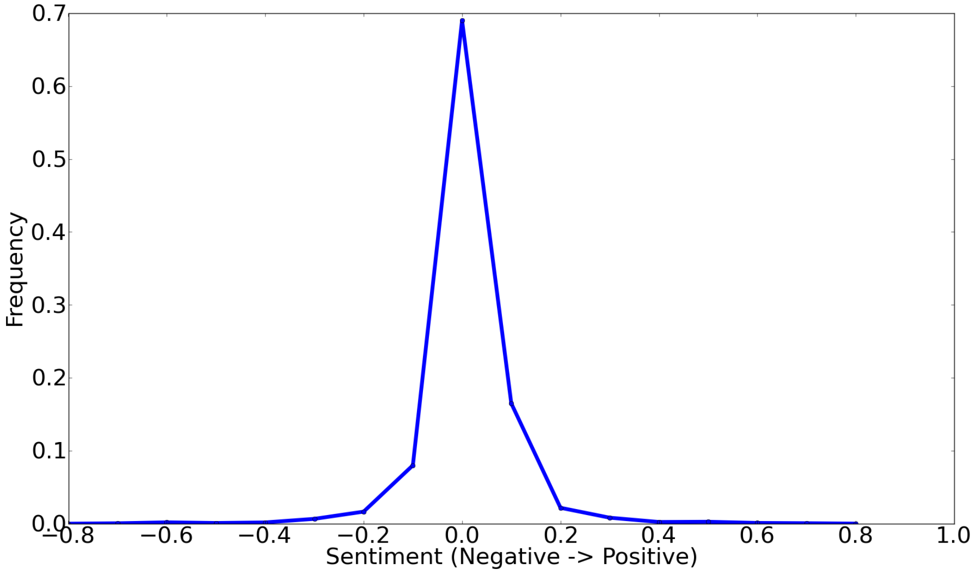


Figure 3.16: Frequency of Users in Categories of Sentiment

Table 3.3: Top-10 Significant Terms for Sentiment Category

Sentiment	Top 10 Terms				
$(0.1, 1.0]$	good	like	love	lol	well
	thanks	great	haha	awesome	nice
$(-0.1, 0.1]$	ave	mayor	street	New York	park
	road	blvd	airport	center	home
$[-1.0, -0.1]$	not	hate	bad	f**k	s**t
	damn	wrong	hell	stupid	hiv

of the top neutral terms are likely to be extracted from the auto-generated check-ins. In the two categories with non-neutral sentiment, we can clearly see typical words which indicate strong positive and negative sentiment.

However, when we filter the top-100 most positive and most negative terms to only consider location-related terms, we find that there are no location-specific positive terms, but there are many location-specific negative terms. Examples of the

Table 3.4: Top-20 Location Terms with Negative Sentiment

MTA	Jersey	Redmond	Memphis
Winooski	Ridgewood	Toronto	Greece
Chicago	Cleveland	Calgary	Scottsdale
Beaumont	Petersburg	Ashburn	Buffalo
Richmond	Montreal	Durham	Eugene

words are listed in Table 3.4. On further inspection of the messages containing these words, we can clearly see the strong negative sentiment associated to the content. For example, when people talk about "MTA", they complain a lot about price increases of MTA's tickets, and its poor service (e.g., "Ticket to the country home has increased by $3. NJTransit is worse than the MTA! (@ New York Penn Station w/ 23 others)", and "I know the MTA is a disaster but 2 of 4 machines being unable to read credit cards at AirTrain station is a new low."). This preliminary analysis indicates that users are more likely to express negative sentiment about location, and that locations and location-related concepts associated with negative sentiment can be automatically identified based on location sharing services.

3.7 Summary

In this section, we tackle the challenge of lack of understanding of the geo-social footprints from location sharing services, by providing a large-scale quantitative analysis and modeling of over 22 million check-ins of location sharing service users. Concretely, three of our main observations are: (i) LSS users follow simple reproducible patterns; (ii) Social status, in addition to geographic and economic factors, is coupled with mobility; and (iii) Content and sentiment-based analysis of posts can reveal heretofore unobserved context between people and locations.

4. PUBLIC CHECK-INS VERSUS PRIVATE QUERIES: MEASURING AND EVALUATING SPATIAL PREFERENCE*

In this section, we resolve the challenge of lack of understanding of whether publicly-shared geo-social data help complement (and replace, in some cases) privately held location information. Specifically, we investigate the viability of new publicly-available geospatial information to capture spatial preference. In the experiments, we compare a set of 35 million publicly shared check-ins voluntarily generated by users of a popular location sharing service with a set of over 400 million private query logs recorded by a commercial hotel search engine.

4.1 Introduction

Social scientists and geographers have long been interested in modeling the linkages and flows between locations for better understanding a variety of geo-spatial issues including: why and how migration flows among countries, regions, and cities; to model commerce flows and explain trade relations among trading partners; to design more efficient roadways and traffic forecasting; to develop epidemiological models of disease spread; and so forth. This *spatial interaction* is a cornerstone of geographic theory, "encompassing any movement over space that results from a human process" [50]. Traditional methods for modeling these flows and the *spatial preference* of users in one location for another location have typically relied on expensive and hard-to-maintain data sources, like the 10-year US Census, which collects massive statistics about the connections between people and between cities in the

*Reprinted with permission from "Public Checkins versus Private Queries: Measuring and Evaluating Spatial Preference" by James Caverlee, Zhiyuan Cheng, Wai Gen Yee, Roger Liew, and Yuan Liang, 2012. *in 5th ACM SIGSPATIAL International Workshop on Location-Based Social Networks.* Copyright 2012 by ACM.

United States.

As a point of excitement, the rise of the web over the past ~20 years has seen a commensurate rise in the low-cost collection of implicit linkages and flows among users and locations. For example, millions of people share their location information passively while using on-line services like video streaming services (e.g., Amazon Instant Video, and Netflix), search engines (e.g., Google, Bing), e-commerce sites (e.g., eBay, and Amazon), and travel planning sites (e.g., Orbitz, Expedia, and Priceline). By tracking IP addresses, plaintext queries, and other location identifiers, these proprietary services have been harvesting huge databases of spatial interaction. For example, by aggregating user search and purchase decisions, Amazon can identify the interest level of users in one location for another location (e.g., more customers in California are buying Texas guidebooks, which may be an early indicator of future migration). However, the excitement over these sources of spatial interaction must be tempered by the proprietary nature of the data.

Fortunately, the past few years have seen the widespread *voluntary* sharing of location information by users of location-sharing services like Twitter, Foursquare, and Google Local. This voluntary sharing about their life, interests, and footprints in real-time provides unprecedented opportunities to study people in different regions, and the connections between people and places. In comparison with expensive, proprietary, and often times unavailable resources, this publicly-shared data offers the promise of new methods appealing not only to geographers and social scientists, but to computational researchers and practitioners seeking to create and improve location-based recommendation systems, travel planners, search engines, and other emerging mobile applications.

Hence, in this section, we investigate the viability of new publicly-available geospatial information to capture spatial preference. Concretely, we explore the spatial preference of users from two large-scale datasets: a set of private query logs for hotels automatically recorded by a commercial on-line hotel search engine (Orbitz), and a set of publicly available check-ins voluntarily generated by users from a typical location sharing service (Gowalla). The check-in data includes over 35 million check-ins from 1.2 million users from Gowalla. The hotel query log data includes all the queries and bookings for hotels from the hotel search engine in 2011, which in total includes over 400 million records from over 20 million unique IPs. We explore in this section the commonalities and the differences between these two sources of spatial preference – generated by different user bases with fundamentally different intentions.

Concretely, this section makes three contributions:

- First, we model the spatial preference of users across both datasets and measure the relative geo-spatial "footprint" of different locations via three localness metrics: the mean contribution distance, the radius of gyration, and the city locality. We find that though the absolute values of these metrics differ across datasets, the relative values are surprisingly consistent.
- Second, we develop a PageRank-like method for identifying spatially significant locations based on the spatial preference of users. Through a random walk over the spatial preference graph linking locations, we find that both datasets reveal similar significant locations.
- Third, we investigate the potential of mining related clusters of locations from both datasets based on the spatial preferences of users. In a comparison against a ground truth of 800 hand-curated lists of related cities, we find similar performance across both public and private datasets.

These results indicate the viability of publicly shared location information via check-ins to complement (and replace, in some cases), privately held location information such as that in proprietary query logs. The potential of publicly shared location information serving as a substitute for privately held information could provide new avenues of research for social scientists, geographers, as well as computer scientists interested in the geo-spatial flows of ideas, memes, and geo-targeted applications.

4.2 Related Work

Researchers have been investigating the spatial properties of large-scale data for many years. In the context of query logs, there have been several efforts typically targeted at the spatial properties revealed through text-based queries to large search engines. For example, Backstrom et al. [6] introduced a model of spatial variation for analyzing the geographic distribution of queries using Yahoo's query logs. The authors proposes a generative probabilistic model in which each query has a geographic focus on a map (based on an analysis of the IP-address-derived locations of users issuing the query). Gan et al. [44] conduct an analysis of 36 million queries from AOL, and identified typical properties for queries with a geographic intention. In addition, they built a classifier that can accurately classify queries into geographic and non-geographic queries.

With the rise of online social networks, there has been a similar rise in analyzing the spatial patterns revealed. For example, Facebook researchers [7] observed that Facebook users have more local friends than distant friends, and that they can predict a Facebook user's location with high accuracy given the location for the users' friends. McGee et al. [85] investigate the relationship between the strength of the social tie between a pair of friends and the distance between the pair with a set of 6 million geo-coded Twitter users and their social relations. They observed that users with stronger tie strength (reciprocal friendship) are more likely to live near each other than users with weak ties. Hecht et al. [52] study the localness of user generated content in Flicker and Wikipedia, and they observe that the content generated by Flickr users is more local comparing to the content generated by Wikipedia editors. A host of related work has also focused on mining interesting trajectories [136], modeling periodic behaviors and mobility patterns [48, 13], and studying the correlation between people's social relations and their mobility patterns [21]. Others have focused on location recommendation at the point of interest (POI) level based on queries and bookings for hotels [99], and check-ins in location sharing services [128, 129, 107]. In the granularity of city-level, researchers have studied the interaction between cities via on-line social relations [64].

4.3 Data

As the basis of this investigation, we consider two large-scale datasets: a set of private query logs and a set of publicly available check-ins.

4.3.1 Private Spatial Resource: Query Logs

The hotel query log data includes a large set of both queries and bookings for hotels randomly sampled from a commercial on-line hotel search engine – Orbitz. The dataset includes over 400 million records, from over 20 million unique IPs all over the world. Each query (or booking) includes an IP address which can be translated to a city-level location where the query (or booking) is issued. We call this the origin location. Each query (or booking) also contains another city-level location indicating the destination (i.e., the city where the queried hotel is located).

To focus on legitimate users of the Orbitz search engine, we filter out IP addresses accounting for an anomalous number of searches (greater than 2,000 queries each). For example, several thousand IPs generate from thousands to millions of queries

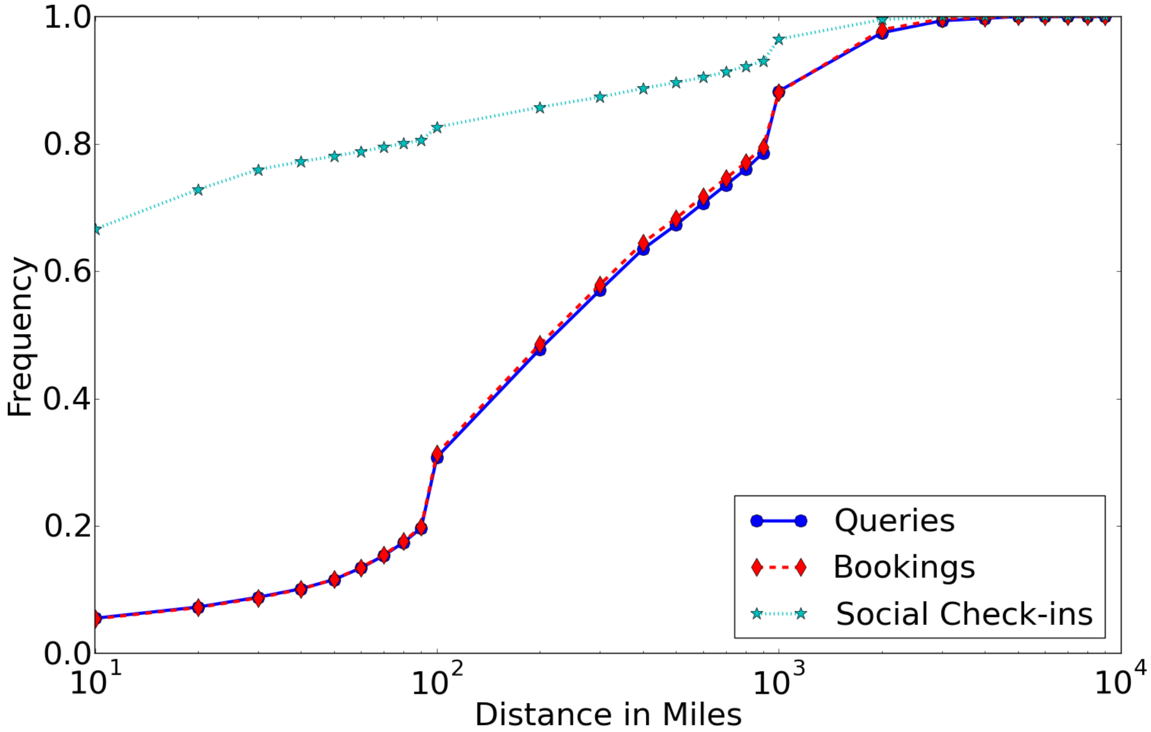


Figure 4.1: Distance versus Frequency: Check-ins tend to be more local; 80% of all check-ins are within 100 miles of a user's home location. In contrast, query (and booking) locations are more distant; only 25% are within 100 miles of a user's home location.

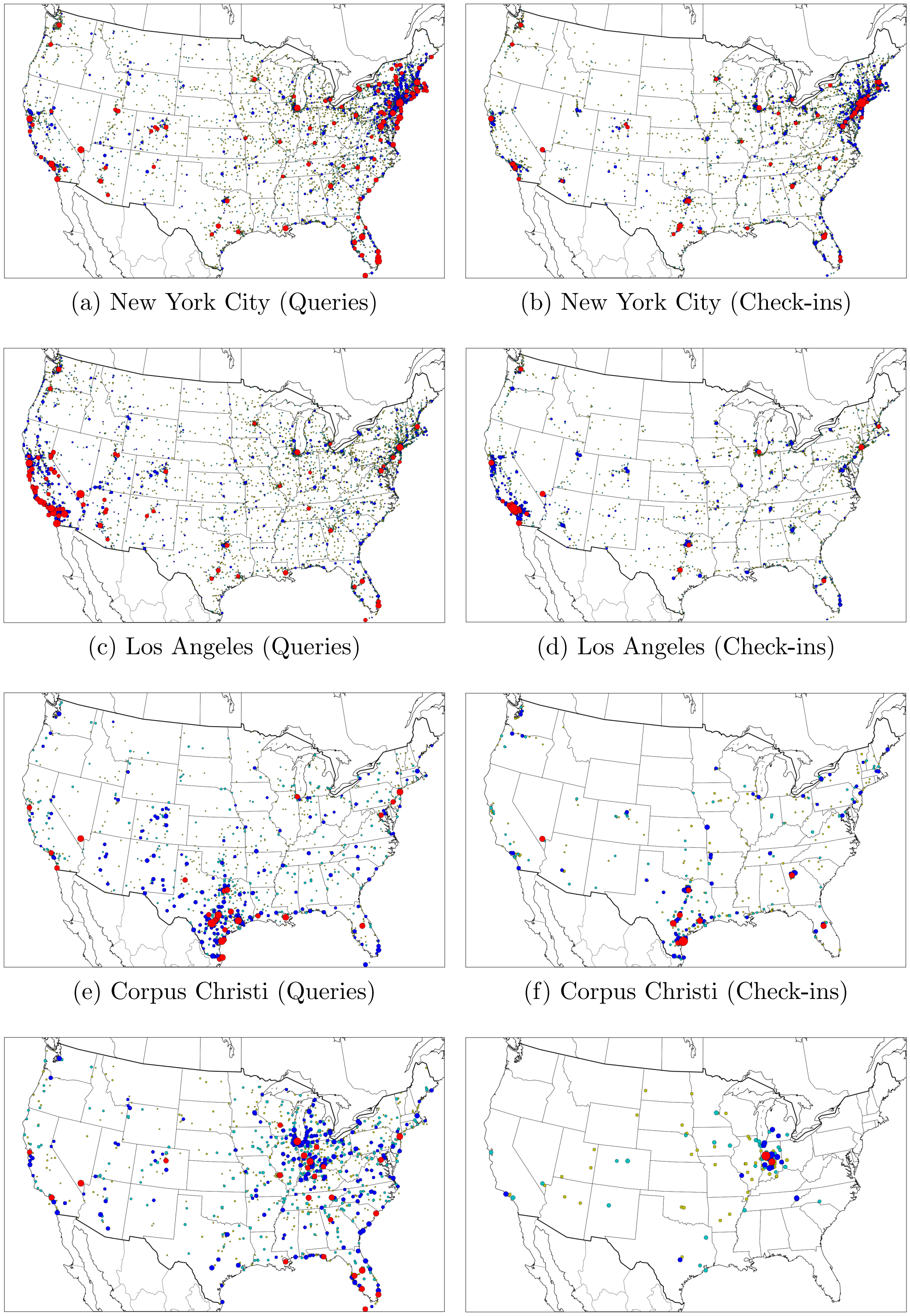
each; most likely, these are search engine crawlers or bots from other travel search engines). Additionally, we focus on queries (and bookings) originating from the Continental United States. Considering each unique IP as a unique user, we consider the corresponding city-level location for the IP as the home location for the user, resulting in **69 million queries and 1.1 million bookings**.

4.3.2 Public Spatial Resource: Check-ins

The check-in dataset includes over 35 million check-ins from about 1.2 million users from Gowalla, a popular location-sharing service. Each of the check-ins includes a fine granular point of interest (POI) location (i.e., where the check-in happened), a timestamp (i.e., when the check-in happened), and a piece of short text (i.e., what the check-in is about). Each check-in's POI location links to a particular city, which allows us to group the check-ins into city-level locations. For each user, we simply consider the city which has the most check-ins from the user as the home location. Similar to the query log data, the check-in data also reveals each user's interest in other "destinations", in this case by considering check-in locations outside of the user's home location. For example, a user from Los Angeles who checks-in in New York City indicates that user's interest in New York. As in the case of the query log data, we focus only on locations within the Continental United States and we filter out users with fewer than 20 check-ins each. The filtering leaves a set of almost 70,000 users and over **15 million check-ins** from the users.

4.3.3 Private versus Public

These two resources – one private and one public – are naturally quite different. Users of these two services vary in their demographics since location sharing service users tend to be young with access to a mobile device, while hotel search engine users are more often representative of the general public with access to a desktop computer. And of course, users of these two services have fundamentally different goals. Hotel queries reflect a user's future intent; check-ins reveal a user's current physical movement. Hotel query logs are more likely to reveal long-distance travel intentions, whereas check-ins are typically a more local phenomenon reflecting a user's interest in local restaurants, bars, and stores [18]. Users of location sharing services are also intentionally sharing their location information, whereas users of search engines are not consciously sharing their location with others (though these search engines may log and analyze the user's queries, IP address, and other location-revealing artifacts). With these many differences in mind, we next turn to an investigation of



(a) New York City (Queries)

(b) New York City (Check-ins)

(c) Los Angeles (Queries)

(d) Los Angeles (Check-ins)

(e) Corpus Christi (Queries)

(f) Corpus Christi (Check-ins)

(g) West Lafayette (Queries)

(h) West Lafayette (Check-ins)

Figure 4.2: (Color) Spatial Preference for Example Cities. Figures in the left-column are derived from private query logs. Figures in the right-column are derived from public check-ins. The color and size of the dots indicate the intensity of the spatial preference from the origin to the destination: top 2% (red); 2-20% (blue); 20-50% (cyan); and the bottom 50% (yellow).

the spatial preference embedded in these two sources and whether we can find any commonalities between them. Finding such commonalities could demonstrate the potential of publicly shared location information serving as a substitute for privately held information.

4.4 Exploring Spatial Preference

We begin our investigation by exploring the spatial preference revealed through both datasets. We model the spatial preference and measure the relative geo-spatial "footprint" of different locations via three localness metrics: mean contribution distance, radius of gyration, and city locality.

4.4.1 Preliminaries

Each query (or booking) in the private dataset and each check-in in the public dataset reveals a bidirectional relationship between an origin location and a destination location. In the case of queries (or bookings) the origin is the city-level location of the user issuing the query; the destination is the city-level location of the hotel. In the case of the check-ins, the origin is the user's home location (which we define as the city with the most check-ins by the user); the destination is the city-level location of the current check-in.

To start with, we are interested in investigating the basic properties of these origin-destination relationships. For each set of queries, bookings, and check-ins, we bucket all the distances between origins and destinations into groups. Figure 4.1 plots the cumulative frequency of the pairs of origin and destination bucketed into groups of distance. The patterns of the bookings and the queries are almost identical to each other, with over 5% of the queries (bookings) for hotels within 10 miles, and about 30% within 100 miles. On the other hand, the check-ins are much more local comparing to the hotel queries (bookings). Over 65% of the check-ins are within 10 miles to the users' home locations, and over 80% are within 100 miles. This difference is our first sign that these two resources reflect fundamentally different usages: that people use hotel search engines to look for hotels to stay during their business trips or vacations, and people use location sharing services to share the real-time status of their daily activities.

4.4.2 Spatial Preference

Given pairs of origin location and destination location extracted from queries (bookings) and check-ins, we quantify the *spatial preferences* for each of the cities with a spatial preference probabilistic distribution. Spatial preference is intended to reflect the aggregate interest level of users in an origin location for a particular destination location.

Spatial Preference: Let l_i be an origin location and let l_j be a destination location. Let $S(l_i)$ be a set of all pairs of origin-destination records in the dataset that originate from location l_i, and let $S(l_i, l_j)$ be a set which includes all pairs of origin-destination records that originate from location l_i with a destination in l_j. Then the spatial preference for location l_i toward location l_j is:

$$p(l_i, l_j) = \frac{|S(l_i, l_j)|}{|S(l_i)|}$$

Example: For example, suppose we have 10 total records (either from the query data or the check-in data) with an origin location of A. Of these, there are three occurrences of ¡A, A¿, two occurrences of ¡A, B¿, and five occurrences of ¡A, C¿. Then, the spatial preferences for location A toward locations A, B, and C are: $p(A, A) = \frac{3}{10} = 0.3$, $p(A, B) = \frac{2}{10} = 0.2$, and $p(A, C) = \frac{5}{10} = 0.5$. Hence, users in location A have the strongest preference for location C, and the weakest preference for location B.

Table 4.1: Average Value for Cities' Localness Metrics

Localness Metric	MCD (miles)	R_g (miles)	CL
Queries	869.346	549.904	0.560
Bookings	809.456	522.644	0.569
Check-ins	380.121	134.477	0.614

Given the definition of spatial preference, we map the spatial preference originating from four cities across both the private query data and the public check-in data. Figure 4.2 highlights the spatial preference of New York City, Los Angeles, Corpus Christi (Texas), and West Lafayette (Indiana). In each of the figures, the color and size of the dots indicate the intensity of the spatial preference from the origin to the destination: red indicates the top 2% most preferred cities; blue indicates the top 2% to 20%; cyan indicates the top 20% to 50%; and yellow indicates the bottom 50%.

As we observe in the figures, the private query data is much denser compared to the check-in data. This is partially an artifact of the data collection limits we faced but is also a reflection of the relative density of these two sources – query logs are inherently a much larger potential collection than are check-ins. Even with this difference in density, we note the relative similarity of the spatial preferences measured across source. People from New York are most interested in the northeast corridor; people from Los Angeles are most interested in the west coast; similar observations can be made for the much smaller locations of Corpus Christi and West Lafayette.

Additionally, we observe that queries balance their locality with many distant locations. For example, Figure 4.2a shows that New Yorkers have many queries for hotels in the New England area, but they are also interested to travel to the Florida and to the west coast. Similarly, Figure 4.2c also shows a a balance between local queries and for more distant ones. In comparison, the check-in data – though of a national scale for both New York and Los Angeles – is much more local (further confirming the relative localness of check-ins versus queries in Figure 4.1). Queries for hotels are relatively more local for the two smaller cities, as we can see in Figure 4.2e and Figure 4.2g. In comparison, the check-in spatial preferences are much sparser and more focused around the origin location.

4.4.3 Comparing Localness

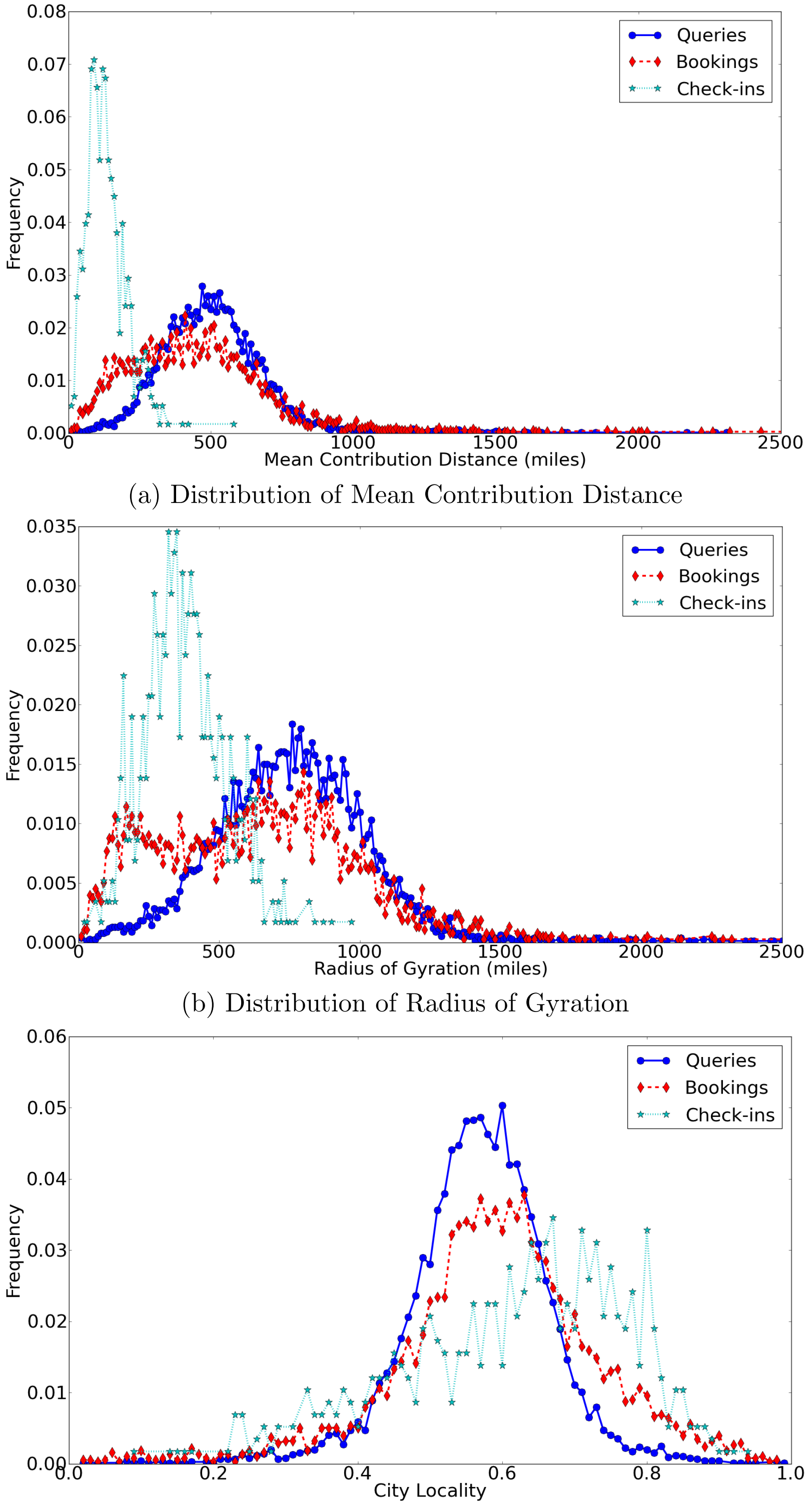
Given the spatial preference probabilistic distribution for a specific location, we can describe each location by measuring its *localness*. The goal of such a localness measure is to encode the entire distribution of spatial preferences into a single summary metric. By evaluating each location, we can directly compare the localness of locations as described by private query logs and by public check-ins. Toward this goal, we adopt three complementary measures of localness:

4.4.3.1 Mean Contribution Distance (MCD)

Proposed by Hecht et al. [52], the MCD measures the weighted average of the distances between an origin location and multiple target locations:

$$MCD(l_i) = \Sigma_{l_j \in S} \left(\frac{d(l_i, l_j) * |S(l_i, l_j)|}{|S(l_i)|} \right)$$

where S includes all locations of interest and $d(l_i, l_j)$ denotes the distance between the origin location l_i and a target location l_j. A small value indicates strong localness for a city; most users in the origin location either query for or check-in to nearby locations. A large value indicates more global interest; users either query for or check-in to distant locations.



(a) Distribution of Mean Contribution Distance

(b) Distribution of Radius of Gyration

(c) Distribution of City Locality

Figure 4.3: Distribution of Localness Metrics

Table 4.2: Values of Localness Metrics for Example Cities

Localness Metric	MCD (miles)		
City Name	Queries	Bookings	Check-ins
New York City	812.384	932.113	310.563
Los Angeles	627.814	619.859	174.568
Corpus Christi, TX	435.364	356.841	172.819
West Lafayette, IN	599.479	543.760	121.559
Localness Metric	R_g (miles)		
City Name	Queries	Bookings	Check-ins
New York City	1278.979	1360.094	747.878
Los Angeles	1056.731	1017.116	541.458
Corpus Christi, TX	693.989	565.083	432.333
West Lafayette, IN	887.397	818.425	282.017
Localness Metric	CL		
City Name	Queries	Bookings	Check-ins
New York City	0.418	0.389	0.334
Los Angeles	0.551	0.552	0.637
Corpus Christi, TX	0.581	0.618	0.231
West Lafayette, IN	0.510	0.539	0.476

4.4.3.2 Radius of Gyration (r_g)

Adopted for location analysis by Gonzalez et al. [48], the r_g measures the standard deviation of distances between an origin location and target locations:

$$r_g(l_i) = \sqrt{\frac{1}{|S(l_i)|} \sum_{l_j \in S} (d(l_i, l_j))^2 * |S(l_i, l_j)|}$$

In essence, the radius of gyration measures both how frequently and how far people from the origin travel. A low r_g typically indicates a location whose residents travel mainly locally, while a high radius of gyration indicates a location with many long-distance travelers.

4.4.3.3 City Locality (CL)

The third measure of "localness" is city locality, proposed by Scellato et al. [103]. The city locality for city (l_i) is formally defined as:

$$CL(l_i) = \frac{1}{|S(l_i)|} * \sum_{l_j \in S(l_i)} |S(l_i, l_j)| * e^{-d(l_i,l_j)/\beta}$$

where β is a scaling factor used to normalize the values of localities so that city localities can be compared using different data and geographic sizes. The city locality is always normalized between 0 and 1. A city with high localness has a higher value of city locality. In practice, the scaling factor β is picked as the mean distance between all the pairs of spatial preference between different cities.

Provided the three localness metrics above, we compare the localness between different cities via their localness metrics. To calculate the localness metrics for each of the cities, we firstly filter out cities without dense data. Specifically, for queries (or bookings), cities with fewer than 1000 queries are filtered out. Similarly, for check-ins, cities with fewer than 1000 check-ins are filtered out. Based on the remaining cities, we calculate each of the three localness measures across queries, bookings, and check-ins.

Table 4.1 shows the average values of the three localness metrics for cities in the three datasets. We see that the private queries (and bookings) naturally reveal a larger scope of interest as compared to public check-ins. The MCD is around 400 to 500 miles greater; the radius of gyration is around 400 miles greater, and the city locality measure is lower (indicating less localness in comparison). Intuitively, it seems reasonable that check-ins are much more local since they are more constrained by physical mobility (e.g., I have to travel to the location, then reveal my location).

As a side note, we see that queries are even less local than bookings, suggesting the exploratory possibility of querying, versus the reality of actually booking a hotel (e.g., it's fun to consider far-flung trips, but in actuality we tend to book more reasonable destinations).

Further confirming this intuition, we show in Figure 4.3, the complete distribution for each of the three localness measures across the private queries (and bookings) versus the public check-ins. We see that the distributions are approximately Gaussian with the check-in distribution resulting in smaller mean contribution distance and smaller radius of gyration, relative to the others. The city locality for check-ins is also skewed more rightward, again conveying the more localness of the check-in data. Connected to the earlier side note, we can see that the bookings are more local than queries based on their distributions.

Finally, we can revisit our four example cities – New York City, Los Angeles, Corpus Christi, and West Lafayette – in terms of the three localness metrics. As shown in Table 4.2, comparing to an average city, people from New York City really travel to a lot of distant cities even farther than the places they searched for. For Los Angeles, the bookings are only slightly more local compared to the queries, while the differences between bookings and queries for Corpus Christi and West Lafayette are even larger than the average gap between queries and bookings. Here our hypothesis is that the gap between localness of queries, and bookings for a particular city might be correlated with the city's demographic information such as population and economy, plus impacted by geographic constraints (e.g., Los Angeles is on the ocean, whereas West Lafayette is in the middle of the country).

4.4.4 Summary

So far, we have modeled the spatial preference of users across both datasets and measured the relative geo-spatial "footprint" of different locations via their mean contribution distance, the radius of gyration, and the city locality. We have observed that the private queries (and bookings) are less local than the public check-ins, which casts doubt on the possibility of publicly shared location information serving as a substitute for privately held information. On an encouraging note, though, we have seen that the relative localness values are surprisingly consistent. Continuing this exploration of the spatial preference, we next turn to two studies designed to leverage spatial preference:

- In the first study, we develop a PageRank-like random walk for identifying spatially significant locations based on the spatial preference of users. Do we find that – in spite of their fundamental differences – that the two datasets reveal similar significant locations?
- In the second study, we investigate the potential of mining related clusters of locations from both datasets based on the spatial preferences of users. Do we find comparable performance across datasets? Or does one perform significantly better than the other?

4.5 Study 1: Spatial Impact

In this section, we explore the possibility of aggregating spatial preference information from multiple locations to provide a global perspective on the most "impactful" locations. Automatically deriving the significant locations from a location dataset is an important problem, and one that has potential applications in urban planning (e.g., what neighborhoods are highly-preferred and potentially facing an influx of new residents?), in location-based advertising (e.g., what points-of-interest

Table 4.3: Examples of Impact Metrics

Impact Metric	ImpactRank		
City Name	Queries	Bookings	Check-ins
New York City	0.035931 (2)	0.017182 (2)	0.010677 (1)
Los Angeles	0.013607 (9)	0.006141 (11)	0.005895 (7)
Corpus Christi, TX	0.001476 (62)	0.001271 (89)	0.000458 (146)
West Lafayette, IN	0.000073 (726)	0.000065 (1628)	0.000121 (535)
Impact Metric	D-ImpactRank		
City Name	Queries	Bookings	Check-ins
New York City	0.038831 (2)	0.019854 (2)	0.016864 (1)
Los Angeles	0.016910 (6)	0.008049 (8)	0.009194 (6)
Corpus Christi, TX	0.001077 (74)	0.001019 (103)	0.000309 (206)
West Lafayette, IN	0.000064 (747)	0.000063 (1575)	0.000101 (534)

are more important for a particular demographic target group?), among many others.

In the following, we formally define two approaches for extracting the significant locations from a location dataset and then we examine the locations identified over the private query dataset and the public check-in dataset.

4.5.1 Two Methods for Finding Spatial Impact

For a collection of locations $\mathcal{L}$, our goal is to find an ordering over the locations in $\mathcal{L}$ corresponding the relative spatial impact of locations, so that higher-ranked locations are deemed more significant than lower-ranked locations. While the notion of spatial impact is difficult to evaluate, we examine two approaches grounded in popular web link analysis and assess the orderings generated by each:

4.5.1.1 ImpactRank

The first approach propagates the spatial preference from one location to another, so that in aggregate the locations that are most preferred by locations that are themselves highly-preferred are the most “impactful”. Similar to the PageRank approach for aggregating web links to assign a global importance score to web pages, ImpactRank can be viewed from the perspective of a biased random walker. At each location, the random walker chooses to visit a subsequent location based on the spatial preference of the current location. As in PageRank, the random walker occasionally loses interest in his travels and randomly picks a new starting location. In the limit, this random walk results in a global ordering over all locations based on the time spent by the random walker in each location.

Let S be the set of all locations, and let $S(\to l_i)$ be the set of all locations that express a non-negative spatial preference in l_i, such that $p(l_j, l_i)$ is the spatial preference probability of l_j toward l_i. The ImpactRank for location l_i, denoted by $IR(l_i)$, is then given by:

$$IR(l_i) = d \sum_{l_j \in S(l_i)} IR(l_j) p(l_j, l_i) + (1 - d) \frac{1}{|S|}$$

where d is a damping factor (fixed as 0.85 in our experiments). The ImpactRank scores may be updated iteratively using the power method.

4.5.1.2 D-ImpactRank

ImpactRank measures the impact of a particular location purely based on the spatial preference matrix (which is essentially a transition matrix defined over locations), but without consideration for the actual distance between locations. Our goal is to incorporate this distance so that more distant locations are more rewarded for the same degree of spatial preference than closer locations. For example, suppose the spatial preference from A to B is 0.2 and from A to C is 0.2. If A and B are neighboring cities, but A and C are separated by 100s of miles, then this method can reward city C more since it has attracted interest from farther away. Thus, we extend ImpactRank to D-ImpactRank, by incorporating the physical distance between locations.

Specifically, we calculate the mean contribution distance (MCD) between all pairs of locations. Then for each spatial preference probability from an origin l_i to a destination l_j, we multiply the original probability by a weight of the distance between l_i and l_j divided by the weighted average distance. The distance weighted spatial preference probability $p'(l_j, l_i)$ from l_j to l_i is defined as:

$$p'(l_j, l_i) = p(l_j, l_i) * \frac{dist(l_j, l_i)}{MCD}$$

Then the D-ImpactRank scores are calculated with the distance weighted spatial preference matrix, and the D-ImpactRank scores for cities are expected to reveal both the cities' spatial impacts and the distance of their impacts' reach. The D-ImpactRank for location l_i can then be defined as in ImpactRank but with updated transition probabilities:

$$DIR(l_i) = d \sum_{l_j \in S(l_i)} IR(l_j) p'(l_j, l_i) + (1 - d) \frac{1}{|S|}$$

4.5.2 Measuring Impact

Given the two approaches for measuring spatial impact, we calculate both over the private queries (and bookings) and the public check-ins. We apply each method to the cities in the Continental United States with dense spatial preference data. As before, we filter cities with fewer than 1000 queries (or bookings) and cities with fewer than 1000 check-ins.

We begin by continuing with our earlier example cities – New York City, Los Angeles, Corpus Christi, and West Lafayette – and listing their spatial impact scores and ranks (in parentheses) in Table 4.3. The relative rankings across both approaches and across all three datasets are remarkably consistent with New York ¿ Los Angeles

Table 4.4: Top 10 Most Impactful Cities By ImpactRank

	Queries	*Bookings*	*Check − ins*
No.1	Las Vegas	Las Vegas	New York
No.2	New York	New York	Austin
No.3	Orlando	Chicago	Orlando
No.4	Miami	Orlando	San Francisco
No.5	Chicago	San Diego	Las Vegas
No.6	San Francisco	Miami	Chicago
No.7	San Diego	New Orleans	Los Angeles
No.8	Phoenix	Washington, DC	Bay Lake, FL
No.9	Los Angeles	San Antonio	Anaheim
No.10	Washington, DC	Atlanta	Seattle

Table 4.5: Top 10 Most Impactful Cities By D-ImpactRank

	Queries	*Bookings*	*Check − ins*
No.1	Las Vegas	Las Vegas	New York
No.2	New York	New York	Austin
No.3	Orlando	San Francisco	San Francisco
No.4	San Francisco	Chicago	Orlando
No.5	Miami	San Diego	Las Vegas
No.6	Los Angeles	Seattle	Los Angeles
No.7	San Diego	Los Angeles	Chicago
No.8	Chicago	New Orleans	Seattle
No.9	New Orleans	Washington, DC	Bay Lake, FL
No.10	Washington, DC	Miami	Anaheim

¿ Corpus Christi ¿ West Lafayette. This is an encouraging result and one that fits well with our intuition (especially considering that Corpus Christi is a popular regional tourist destination as compared with the college town of West Lafayette).

We next list the top-10 cities with the highest spatial impact in Table 4.4 and Table 4.5, again considering both approaches and all three datasets. Focusing on Table 4.4, we see that five of the ten cities are common between the public check-in dataset and the private query dataset: New York, Orlando, San Francisco, Chicago,

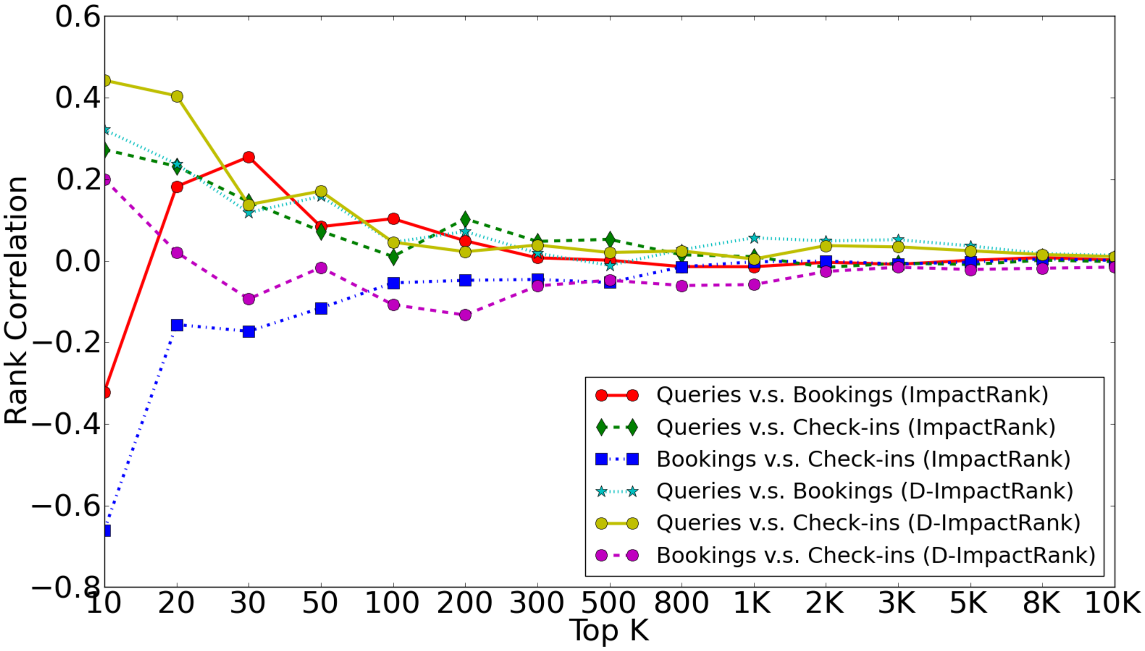


Figure 4.4: (Color) Rank Correlation between List of Most Impactful Cities

and Los Angeles. Note that Austin is the original home of the Gowalla location sharing service and so it receives a large "home field advantage". Bay Lake, Florida is the home of Walt Disney World next to Orlando and so could be considered a sixth similar location across the public and private datasets. Similarly, we see in Table 4.5 comparable rankings for the distance-weighted D-ImpactRank with respect to the original ImpactRank.

Comparing between ImpactRank and D-ImpactRank for only the top-10 reveals little difference. Hence, we next measure the rank correlation across approaches using Spearman's ρ, which ranges from 1 to -1, with higher values indicating that two ranked lists are in relative agreement. As we can see in Figure 4.4, the rank correlation between approaches and between different datasets varies quite a bit. The series of red, green, and blue indicate the rank correlations between lists of top-K most impactful cities ranked by their ImpactRank scores. The series of cyan, yellow, and magenta indicate the rank correlations between lists of top K most impactful cities ranked by their D-ImpactRank scores. We are encouraged to see that the rank correlation for D-ImpactRank for queries versus check-ins performs very well over the top-20 results (meaning that the top-20 are highly correlated based on these two datasets). For bookings versus check-ins over ImpactRank (in blue), the rank correlation is the worst for K up to 100. At higher values of K, the rank correlation in all cases converges to around 0.0 primarily due to data sparsity at the bottom of the ranked list (leading to essentially random rankings at the bottom of the list).

Based on this experimental study, we find that in some cases both datasets reveal similar significant locations. This result is somewhat surprising considering the key differences between the public check-ins and the private queries, but is encouraging. In our following study, we continue this exploration of the viability of substituting publicly-released data for private data with an examination of extracting similar cities from location datasets.

4.6 Study 2: Finding Similar Cities

In previous sections, we characterized a location by its spatial preference and by the spatial impact derived from aggregating over these spatial preferences. In this section, we examine whether these spatial characterizations can be used to automatically extract groups of similar locations. Finding related groups of locations has potential impact for optimizing online advertising (e.g., if users in location A click on an ad, then perhaps users in the similar location B will also do so), for improving web search and mobile applications (e.g., a user querying for a nearby tourist destination can be recommended other similar spots), and so forth.

Toward finding similar cities, we first define a ground truth of city similarity, define two metrics for evaluating city similarity, and then measure city similarity

Table 4.6: Performance for Identifying Similar Cities

	Queries		Bookings		Check-ins	
Feature Set	P@10	N@10	P@10	N@10	P@10	N@10
Spatial Preference	25.17%	56.11%	28.06%	59.81%	24.2%	60.1%
Spatial Impact	22.22%	50.43%	24.71%	52.86%	31.2%	64.7%
Spatial Preference + Impact	28.04%	59.42%	28.97%	60.38%	31.6%	65.2%

using a vector space interpretation of spatial preference and spatial impact.

4.6.1 Defining the Ground Truth

What makes two locations similar? While there are many possible answers, we adopt a systematic method for finding relationships among cities by mining 800 expert-curated lists of top cities across particular categories. The data is available from [1] and lists 101 top cities for each category. For example one of the lists includes the top cities with the most people having a Doctorate degree; for this list the top cities are Palo Alto (CA), Bethesda (MD), Brookline (MA), Cambridge (MA), and Davis (CA). From this perspective, these five cities can be considered similar. In this same fashion, we extract the top cities lists for a total of 800 separate city lists. For each pair of cities, we consider their total number of co-occurrences among the top city lists as the similarity between the pair of cities. For example, if two cities co-occur in 400 out of the 800 lists, then their similarity is $\frac{1}{2}$. Cities that never co-occur on a list will have a similarity of 0. In addition, for city l_i, we rank the other cities according to their similarities (co-occurrences in top city lists) with city l_i in descending order.

A similar approach was undertaken in the context of free-text web search engine queries in [106]. Rather than considering spatial preference as in this section, the authors looked for common clues in the text of search engine queries to group related cities. Information revealed through text queries is a strong indicator of similarity (e.g., if many users in two locations are both querying for “molecular biology”, “PhD”, and “grad school”, then there is good evidence of a relationship between locations). In contrast, spatial preference is a less clear indicator of city similarity since only relative interest in other locations is available for comparison.

4.6.2 Approach and Metrics

To find related cities, we apply the standard cosine similarity to vectors based on the spatial preference and the spatial impact of city pairs. That is, for city i and city j, we can represent each city by a vector (e.g., based on the spatial preference probabilities). Cosine similarity is a similarity measurement between the two vectors – in this case, the vectors associated with city i, $\vec{v}_i$, and with city j, $\vec{v}_j$: $cos(\vec{v}_i, \vec{v}_j) = \frac{\vec{v}_i \cdot \vec{v}_j}{|\vec{v}_i||\vec{v}_j|}$.

With this approach and the ground-truth data, we use **Average Precision@10** (P@10) and **Average NDCG@10** (N@10) to evaluate the predicted top similar cities. For each city, we first extract the top K% of the most similar cities to it in the ground truth data as the relevant cities to the given city. Then, we calculate the Precision@10 for the city which measures the percentage of the top 10 predicted similar cities that also belong to the top K% of the relevant set, which can be formally defined as:

$$P@10 = \frac{\sum_{l_i \in S} \frac{|S_{top10}(l_i) \cap S_{top\_k\%\_gt}(l_i)|}{10}}{|S_c|}$$

where S refers to the set of all the cities in the datasets; l_i denotes a specific city; $S_{top10}(l_i)$ denotes the top 10 similar cities of l_i predicted using the similarity metric; and $S_{top\_k\%\_gt}(l_i)$ denotes the top K% similar cities for l_i in the ground-truth data.

A high value of $AvgPrecision$@10 indicates that the location preferences or localness modeled from the data really reveal semantic information for the city, and hence provide hints to find similar cities. Similarly, we apply **Average NDCG@10** to evaluate the performance considering both the precision of the predicted similar cities and the positions of the truly similar cities in the predicted similar city list.

In practice, we extract 10% of the ground truth similar cities for each city as its ground truth relevant cities. To make sure we have dense data for each of the cities, for both the queries and bookings, we only pick the cities in Continental United States with a minimum of 5000 queries from each of the city; and for the check-ins, we only pick the cities in Continental United States with a minimum of 1000 check-ins.

4.6.3 Evaluation

Table 4.6 shows the performance using features of different combinations of spatial preference and spatial impact associated with the private queries (and bookings) and the public check-ins. We additionally consider a combined vector representation that is simply an average of the normalized spatial preference and the normalized spatial impact vectors. Using cosine similarity to calculate the similarity between these three representations of cities, we observe strikingly similar results across the public check-ins and the private queries, as well as fairly stable relative ordering with the combined representation always yielding the best results.

Focusing on precision@10, we see that about 28% of the top-10 predicted similar cities are considered similar (based on the ground truth data) based on the query data, but that about 32% are similar based on the check-in data. Focusing on the average NDCG@10, we see a similar behavior – with the query data yielding a 60% result, but the check-ins performing slightly better with 65%.

Based on this experimental study, we find that across these two fundamentally different datasets, that similar performance may be achieved for automatically identifying groups of related locations. Coupled with the observations in the previous section, this is a second encouraging result considering the key differences between the two datasets.

4.7 Summary

In this section, we have investigated two different sources of spatial preference: a set of private query logs recorded by a commercial hotel search engine and a set of publicly shared check-ins voluntarily generated by users of a popular location sharing service. Although generated by users with fundamentally different intentions, we find common conclusions may be drawn from both data sources, indicating the viability of publicly shared location information to complement (and replace, in some cases), privately held location information. This is especially encouraging since many location preference data sources are expensive, proprietary, and often times unavailable. In contrast, publicly-shared data offers appealing new avenues of research. Since modeling and exploiting spatial preference is critical for geographers, social scientists, as well as computer scientists interested in improving location-based recommendation systems, travel planners, search engines, and other emerging mobile applications, these conclusions are a starting point for further research on the strengths and weaknesses of relying on publicly available datasets.

5. ACTIVITY-DRIVEN LOCAL SEARCH*

In both section 5 and 6, we talk about the new geo-social information systems we propose. In section 5, we first introduce a location-based search system augmented using activity pattern mined from location-sharing services. Concretely, we get started from studying location-based activity patterns (also referred to as traffic patterns) revealed through location sharing services, and find that these activity patterns can identify semantically related locations. Based on this observation, we propose and evaluate a activity-driven location clustering algorithm that can group semantically related locations with high confidence. Through experimental study of 12 million locations from Foursquare, we extend this result through supervised location categorization, wherein traffic patterns can be used to accurately predict the semantic category of uncategorized locations. Based on these results, we show how activity-driven semantic organization of locations may be naturally incorporated into location-based web search.

5.1 Introduction

The emerging check-ins from location sharing services – along with other user-generated descriptors supported by these services like tags, ratings, and comments – have resulted in billions of explicit "geo-semantic" markers that link people, places, and their activities. As these services continue to grow, there are great opportunities for extremely granular temporal and spatial mining of human mobility, as well as new mobile+location-based services, augmented traffic forecasting, and urban planning.

*Reprinted with permission from "Toward Traffic-Driven Location Based Web Search" by Zhiyuan Cheng, James Caverlee, Krishna Kamath, and Kyumin Lee, 2011. *in Proceedings of the 20th ACM International Conference on Information and Knowledge Management.* Copyright 2011 by ACM.

In this section, we propose one direction in which location-sharing services may have strong impact – in augmenting traditional location-based web search. Location-based web search (also known as local search) has drawn intensive attention in both industry (Google Maps, Yelp, Yahoo! Local, and Yellow Pages), and the academic community (e.g., [15, 32, 84]). Nearly all the current location-based search systems typically provide rankings for nearby venues based on a user's query and current location. For example, a local search for "coffee" may return a map and an associated ranked list of nearby coffee shops and coffee bean wholesalers. Some of the factors that location-based search engines use for ranking venues in response to a query include: (i) the distance between the user and the target venue; (ii) category analysis of venues (e.g., to group all coffeeshops in a pre-processing step); (iii) the overall ratings for the venue (which are often available for commercial places of business like restaurants); (iv) query and click popularity of the venue's associated web page (e.g., Starbucks may be considered more popular from its web presence than a local coffee shop); (v) reputation of the location via PageRank-style link analysis of the web graph; and (vi) content-based relevance between the query and the location's description (e.g., via information retrieval similarity between the query and a summary of the venue on Yelp or the content from a location's web presence).

In many ways analogous to how clickstreams [56, 123, 23] have been successfully incorporated into traditional search systems based on content similarity [101] and link analysis [62] by connecting real-world user actions (clicks) to relevance, this section proposes that the *temporal dynamics embedded in the check-ins* from location sharing services have great potential to augment traditional location-based search systems by connecting real-world actions (check-ins) to relevance. To illustrate:

- Mike wants to make a reservation for a tennis court on Saturday afternoon so that he can teach his son to play tennis without being disturbed by nearby players. Hence, a local search for “tennis courts” could be augmented with the temporal dynamics mined from location-sharing services to indicate which courts are at off-peak times in terms of player traffic.

- Tina and her friends are going to celebrate their graduation on a Thursday evening and are looking for late-night hot spots. Which local bars are at-peak in terms of traffic? Or will be peaking by the time Tina and her friends arrive?
- John plays a lot of basketball. He usually goes to Williams Park during Wednesday early evening, and Saturday afternoon, which are both free time for him and peak times for other players to get together and play basketball. Suppose John moves to a new neighborhood and wants to find places nearby that have similar traffic patterns, so that he can meet new friends there and play basketball. A traffic-driven location-based search can also easily handle this kind of queries by returning semantically correlated venues with similar traffic patterns.

In all three cases, factors traditionally considered for loca- tion-based web search – like distance, overall venue reputation and popularity – are less important than fine-grained temporal dynamics of the traffic patterns of the target venues. Hence, there is an opportunity to augment these traditional approaches with real-world user actions revealed through location sharing services.

In this section, we propose to study the potential and viability of mining traffic patterns revealed through location sharing services to augment traditional location-based search. As a first step, we propose to model each venue by a *traffic pattern* – essentially a frequency function corresponding to each venue. Two essential and open questions are (i) whether such a model, as compared to traditional content-based and popularity-based models of location-based search, encodes semantically meaningful information; and (ii) whether there is wide enough coverage of location sharing services to support large-scale application of traffic patterns to location-based search. With these questions in mind, this section makes the following contributions:

- First, we present in this section a large-scale study of every venue in Foursquare, totaling 12 million unique venues annotated by users of Foursquare via check-ins. Based on this study, we propose and evaluate a *traffic pattern*-based model of venues through an investigation of the location-based traffic patterns mined from 22 million check-ins from Foursquare and other location sharing services.
- Second, we propose a measure of semantic correlation across venues for organizing venues according to the traffic patterns revealed through location sharing services. Based on this measure, we propose and evaluate a traffic-driven location clustering algorithm that can group semantically related locations with a best-effort performance of F1-Measure 0.675, and Purity 0.764, a critical function for a location-based search engine.
- Third, we observe significant sparsity of check-in data for venues on the "long tail", and so we propose and evaluate a traffic pattern-driven approach for supervised location categorization, wherein traffic patterns can be used to accurately predict the semantic category of uncategorized locations with a F1-Measure of almost 0.8.
- Finally, based on these results, we show how traffic-driven semantic organization of locations may be naturally incorporated into location-based web search through two example scenarios.

5.2 Related Work

Increasing focus has been put on location sharing services in recent years. Ye et al. [128] proposed friend-based collaborative filtering algorithms to recommend locations utilizing a dataset scraped from Foursquare; Lindqvist et al. [75] analyzed how and why people use location sharing services; and Noulas et al. [88] analyzed user check-in dynamics, and user activities in location sharing services. Compared to these previous studies, this section focuses on analyzing the temporal traffic patterns revealed from location sharing services and how the traffic patterns can be utilized to enhance traditional location-based search.

Several studies have analyzed the temporal dynamics of on-line social networks and other web corpuses. Golder et al. [47] explored the temporal dynamics associated with on-line social tagging activities. Researchers in [63] studied how queries, their associated documents, and the query intent change over time by analyzing query log data. Temporal evidence was incorporated into models of semantic relatedness for words in [95]. Yang et al. [125] proposed a clustering algorithm that groups temporal patterns associated with online content, and studied how the popularity of the content grows and fades over time. A temporal correlation measure was introduced and applied to study semantic similarity between queries by Chien et al. [20].

Related to our temporal model of traffic patterns, in terms of time series data analysis, Fu [43] provides comprehensive summary on the existing time series data mining literature including representation, indexing, similarity measure, segmentation, visualization and mining. A numerosity reduction component was proposed by Xi et al. [121] and proved to speed up the best performing classifier of one-nearest-neighbor with Dynamic Time Warping (DTW). [72] provided a survey for techniques in time-series data clustering, and corresponding evaluation metrics.

Location-based web search has drawn intensive attention in both industry (Google Maps, Yelp, Yahoo! Local, and Yellow Pages), and the academic community. Early research efforts (e.g., [15, 32, 84]) mainly focused on the extraction of geographic information from page content and structure. Several studies [44, 4, 102] showed that more than one fifth of the queries in general web search systems were geographical relevant queries. Geotagging and gazetteers have been widely used in [81, 115, 117] to augment location-based web search. Watters and Amoudi [118] proposed a method to assign location coordinates to URLs, and a corresponding framework for location-based ranking of search results.

5.3 Location Sharing Services

In this section, we introduce the location sharing service data, present our sampling strategy, and provide a characterization of the venue data collected from Foursquare.

5.3.1 Sampling Check-in Data

eo start with, firstly we need a set of check-ins. While Foursquare, Facebook Places, and other related location sharing services are rich resources, all restrict access to a user's immediate social circle and hence are unavailable for public sampling. Hence, we adopt a data collection technique that relies on sampling location sharing status updates from the public Twitter feed. While users of location sharing services in general and the subset who choose to advertise their location via Twitter may not be a representative sample, these status updates are inherently public (mitigating concerns over privacy violations that would arise from mining services like Facebook Places) and offer a rich vein of check-in data. Specifically, we monitor Twitter's public streaming API and search API from October 2010 to January 2011, and collected a set of more than 22 million check ins. Worth mentioning, our data is available on-line at `http://infolab.tamu.edu/data/`.

Each check-in contains a fine-granularity location (latitude and longitude) and a timestamp. More than 62% (∼14 million) of the check-ins are associated with a

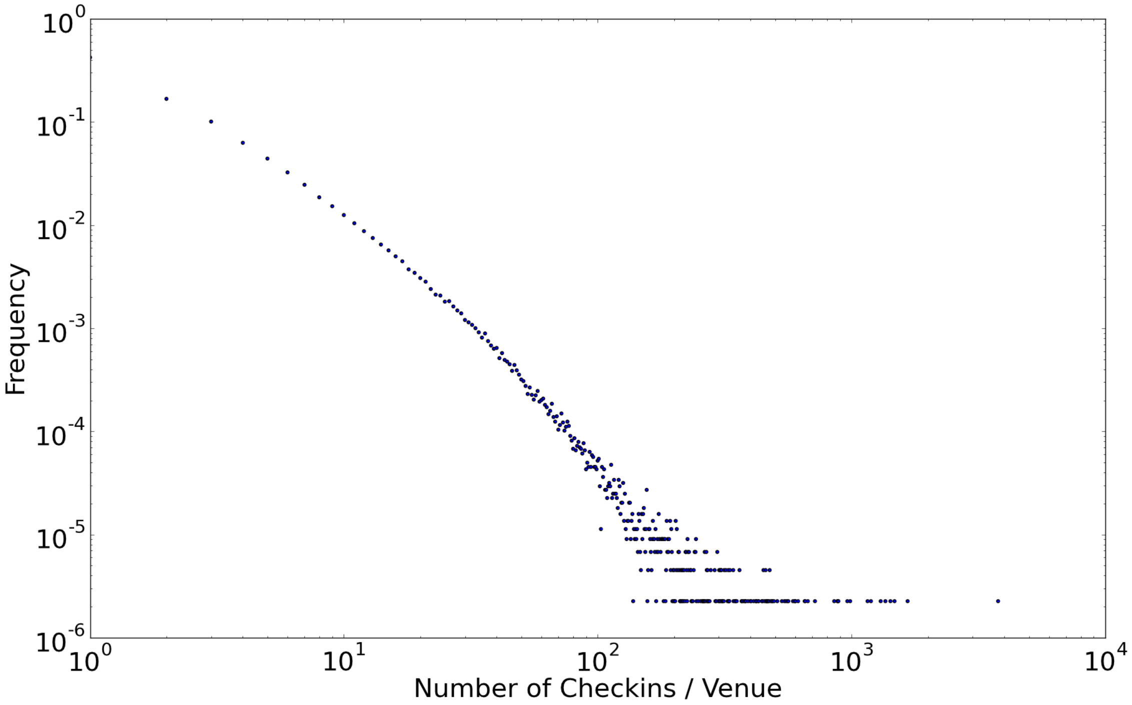


Figure 5.1: Distribution of # of Check-ins Per Venue

venue, and in total 603,796 venues are referenced. Note that since each venue has on average only ∼23 check-ins (with a skewed distribution, where some venues are heavily "checked in" to, but the majority have only a handful of check-ins as plotted in Figure 5.1), we aggregate all check-ins for venues based on the venue name (e.g., grouping all instances of "Starbucks") for the analysis in the rest of the section. Venues in the set may be associated with varying degrees of spatial granularity based on the bounding box linked to the venue – from country to province / state to city to district and finally to points-of-interest. In this section, we mainly focus on the check-ins corresponding with the 515,862 point-of-interest venues, each of which is finely geo-labeled with a latitude and longitude.

5.3.2 Crawling Foursquare Venues

Each venue posted to Twitter has a corresponding "venue page" hosted by Foursquare. To retrieve more information about the venues, we crawled the entire Foursquare-sphere, resulting in nearly 20 million "venue pages" in HTML format. Based on our best-effort parser, we successfully parse 12,677,314 html pages of venues. Specifically, each venue is stored as the tuple *venue(venueID)* = {*venueID, name, latitude, longitude, address, city, region, postal_code, categories, tags*} An example tuple of a venue is: *venue(877)* = {*877, "once upon a tart", 40.7267, -74.0019, "135 sullivan st", "new york", "ny", 10012, "sandwiches, salad, bakery", "salads, roast pork sandwich, strawberry lemonade ginger iced tea, tarts, desserts"*}.

5.3.2.1 Venue Characterization

Among the 12 million venues, 56.4% (i.e., 7,147,755) of the venues are voluntarily assigned by users of Foursquare at least a single category. And there are 7,753,274 occurrences of 833 unique categories identified in the dataset. Based on Foursquare's 3-level categorization system, we group the 833 categories into 8 coarse groups: Arts & Entertainment, College & University, Food, Great Outdoors, Home, Work and Other, Nightlife Spot, Shop, and Travel Spot. The distribution of the eight categories is listed in Table 5.1. Among the categorized venues, the category of Home, Work and Other presents almost one third (31.7%) of the venues. Venues in the category of Food (24.3%) and the category of Shop (17.1%) are also popular. The other five categories (Travel Spot, Great Outdoors, Nightlife Spot, Arts & Entertainment, and College & University) possess similar percentages (around 5% for each) in the dataset.

Besides the category information for venues, about 7.8% (i.e., 989,281) of the venues are labeled with at least a single tag. The tags are user-generated keywords

Table 5.1: Distribution of Venue Categories

Category	Percentage	# of Venues
Home, Work and Other	31.7%	2,457,172
Food	24.3%	1,886,875
Shop	17.1%	1,329,185
Travel Spot	7.0%	541,482
Great Outdoors	6.4%	493,635
Nightlife Spot	5.7%	438,400
Arts & Entertainment	4.3%	334,700
College & University	3.5%	271,825

posted by users of the location sharing service. Based on inspection, tags typically contain information such as category of the venue (e.g., coffee, food, and bar); items provided by the venue (e.g., burgers, flu shot, and long island iced tea); features of the venue (e.g., free wifi, 24 hrs, and pet friendly); location of the venue (e.g., houston downtown, bridge street); and users' comments for the venue (e.g., awesome, good deal, and great food). Each of the tagged venues is assigned with an average of 3.37 tags. Different from the categorization system, the tagging feature in Foursquare gives users more freedom to generate appropriate tags. In total, we find 615,457 unique tags that are collectively used a total of 3,329,641 times across all venues.

Together, these user-assigned tags and the top-level categories provide descriptive information about specific venues and provide clues to study the semantic correlation between venues. Recall that one of the key pre-processing steps in location-based search is category analysis of venues – to group together semantically-related venues – but in isolation we can see that the category assignments are fairly sparse (56%) and at a coarse-level; similarly, the tag information is even sparser (8% of all venues), and both tags and categories provide only *descriptive* information about the venues. Our goal in the rest of the section is to consider the traffic-driven temporal patterns revealed through check-ins to augment this semantic grouping based on the real-world behaviors of users of these services.

5.4 Exploring Semantic Correlation between Venues

In this and the following two sections, we begin an exploration of the temporal dynamics of venues as revealed through location sharing services. Given the large-scale chec-kin data, we propose to model venues through traffic-based patterns and seek to answer the following questions:

- Can we measure semantic correlation between venues based on associated traffic patterns?
- Can we cluster venues into semantically correlated groups based on traffic patterns?
- Can we use traffic patterns to accurately predict the semantic category for uncategorized locations?

We begin in this section by defining a traffic pattern and its frequency function. We discuss metrics to measure semantic similarity between traffic patterns, and we apply the temporal correlation measure to quantify the semantic relatedness between traffic patterns. Based on this initial study, we identify semantically correlated groups of venues based on measuring the pairwise temporal correlation between the venues' associated traffic patterns.

5.4.1 Modeling Venues

A **Traffic Pattern** (T Pattern) T for a venue over n time units is defined as the temporal dynamic of check-ins during the time period. It can be measured by its **Frequency Function** F_T formally defined as $F_T = (f_{t_1}, f_{t_2}, ..., f_{t_n})$, in which f_{t_i} is the frequency for time unit t_i over the whole series of T. More specifically, for each

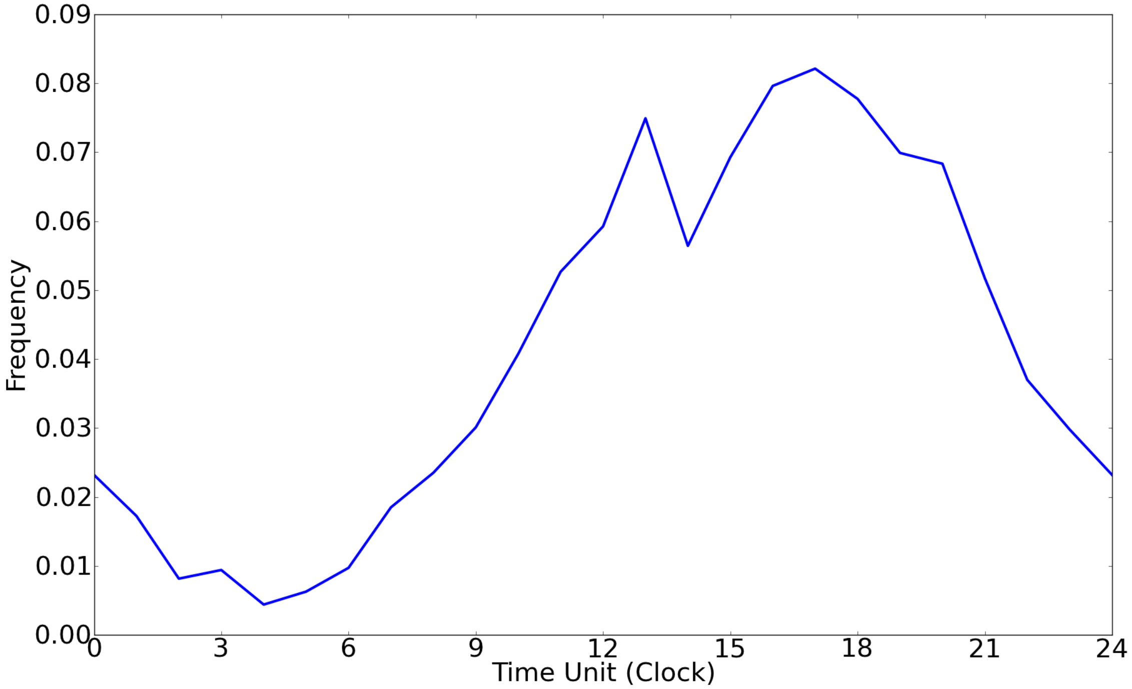


Figure 5.2: Daily Traffic Pattern for Walmart

venue, we generate a daily mean traffic pattern and a weekly mean traffic pattern given the timestamps of check-ins in the venue. The **Daily (Mean) Traffic Pattern** contains 24 time units in which each of the time unit represents an hour in a day. Similarly, the **Weekly (Mean) Traffic Pattern** contains 70 time units in which each unit represents one tenth of a day. Examples of daily traffic pattern and weekly traffic pattern for Walmart are plotted in Figure 5.2 and Figure 5.3 respectively. The daily t pattern shows that customers tend to go shopping in Walmart in the afternoon and early evening, and the weekly t pattern indicates that there is a bigger crowd at Walmart over the weekends than on weekdays.

5.4.2 Temporal Similarity Measures

Given a traffic pattern for a venue, can we identify related venues based solely on this pattern? This is an important step for semantically grouping venues for

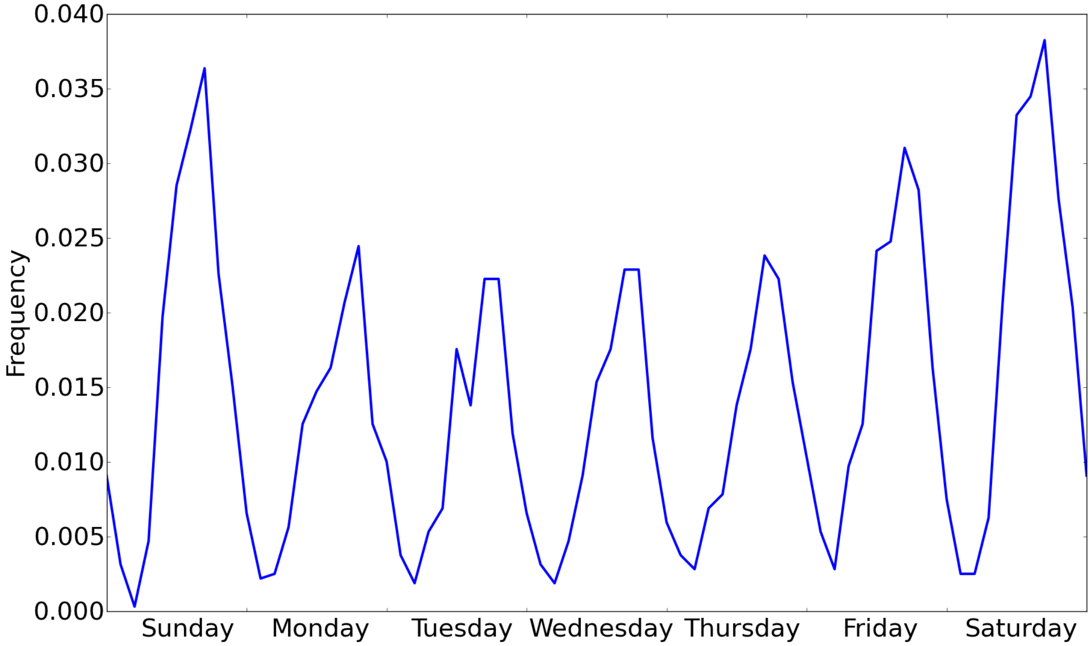


Figure 5.3: Weekly Traffic Pattern for Walmart

improved location-based search. But perhaps traffic patterns do not vary much from venue to venue, meaning that traffic patterns could have only limited impact.

The most straightforward similarity measures for time-series data are Euclidean Distance [36] and its variants based on the common $L_p - norms$ (L_1 – Manhattan Distance, and L_2 – Euclidean Distance). These metrics can be easily implemented and are surprisingly competitive with other complex measures with a large training set. However, these distance measures are sensitive to noise and misalignments in time. Another effective temporal correlation measure is a temporally-grounded variation of the correlation coefficient. Given two traffic patterns Tp, Tq and their frequency functions F_{Tp} and F_{Tq}, the temporal correlation $T_{Corr}(F_{Tp}, F_{Tq})$ between the two traffic patterns Tp and Tq is:

$$T_{Corr}(F_{Tp}, F_{Tq}) = \frac{1}{n}\sum_{i}(\frac{f_{tp_i} - \mu(F_{Tp})}{\sigma(F_{Tp})})(\frac{f_{tq_i} - \mu(F_{Tq})}{\sigma(F_{Tq})})$$

where $\mu(F_{Tp})$, $\mu(F_{Tp})$ are the mean frequencies, and $\sigma(F_{Tp})$, $\sigma(F_{Tq})$ are the standard deviations for the two traffic patterns Tp and Tq. A version of this same metric was shown to be effective by Chien et al. [20] for measuring the similarity of search engine queries by comparing their frequency functions.

5.4.3 Mining Semantic Correlation between Venues

To study the semantic correlation between venues based on their traffic patterns, we sample a set of 271 venues from the check-ins dataset with a criteria of at least 100 branches and 100 check-ins to ensure the density of the traffic patterns. For each of the 271 venues, we retrieve both a mean daily traffic pattern and a mean weekly traffic pattern to capture their traffic. Then, we calculate the pairwise temporal similarity for all the pairs in the 271 venues set. After sorting the pairs of venues based on the descending order of temporal correlation measure, we find quite a few interesting pairs of venues that have obvious semantic correlation. Using the results for calculations based on daily mean traffic patterns only, we show the top-10 similar pairs of venues in Table 5.2. Each pair of venues in the table has obvious semantic correlation: both “Walgreens” and “CVS Pharmacy” are 24 hour pharmacy stores; both “Subway” and “Jason’s Deli” are chain fast food restaurants; and both “Starbucks” and “Caribou Coffee” are coffee shops. The results listed in the table clearly indicate that the traffic pattern of a venue reveals its semantic category, and the temporal correlation between traffic patterns of two venues can help measure the semantic relatedness between venues.

Having all pairwise temporal similarities between venues, we are also interested to see whether we can find inherent groups of venues that belong to the same semantic

Table 5.2: Top Pairs with Highest Temporal Correlation

Pair of Venues	Correlation
Target – Borders	0.949
Walgreens – CVS Pharmacy	0.947
Panda Express – Five Guys Burgers and Fries	0.947
Pizza Hut – California Pizza Kitchen	0.947
Chipotle – Five Guys Burgers and Fries	0.946
Staples – Apple Store	0.946
Target – Barnes & Noble	0.946
Subway – Jason's Deli	0.945
Chili's – Ruby Tuesday	0.944
Starbucks – Caribou Coffee	0.944

category. For example, do all the coffee shops have similar temporal traffic patterns? We model the venues and temporal similarities as vertices and weights for edges in a graph. An edge between two vertices (venues) exists when the temporal correlation between traffic patterns of the two venues exceeds a pre-defined threshold. In this way, a graph modeling the semantic relationship between venues is generated. Instead of focusing on the whole graph itself, we are more interested in the strong connected components in the graph, which are potential candidates for semantic categories of venues.

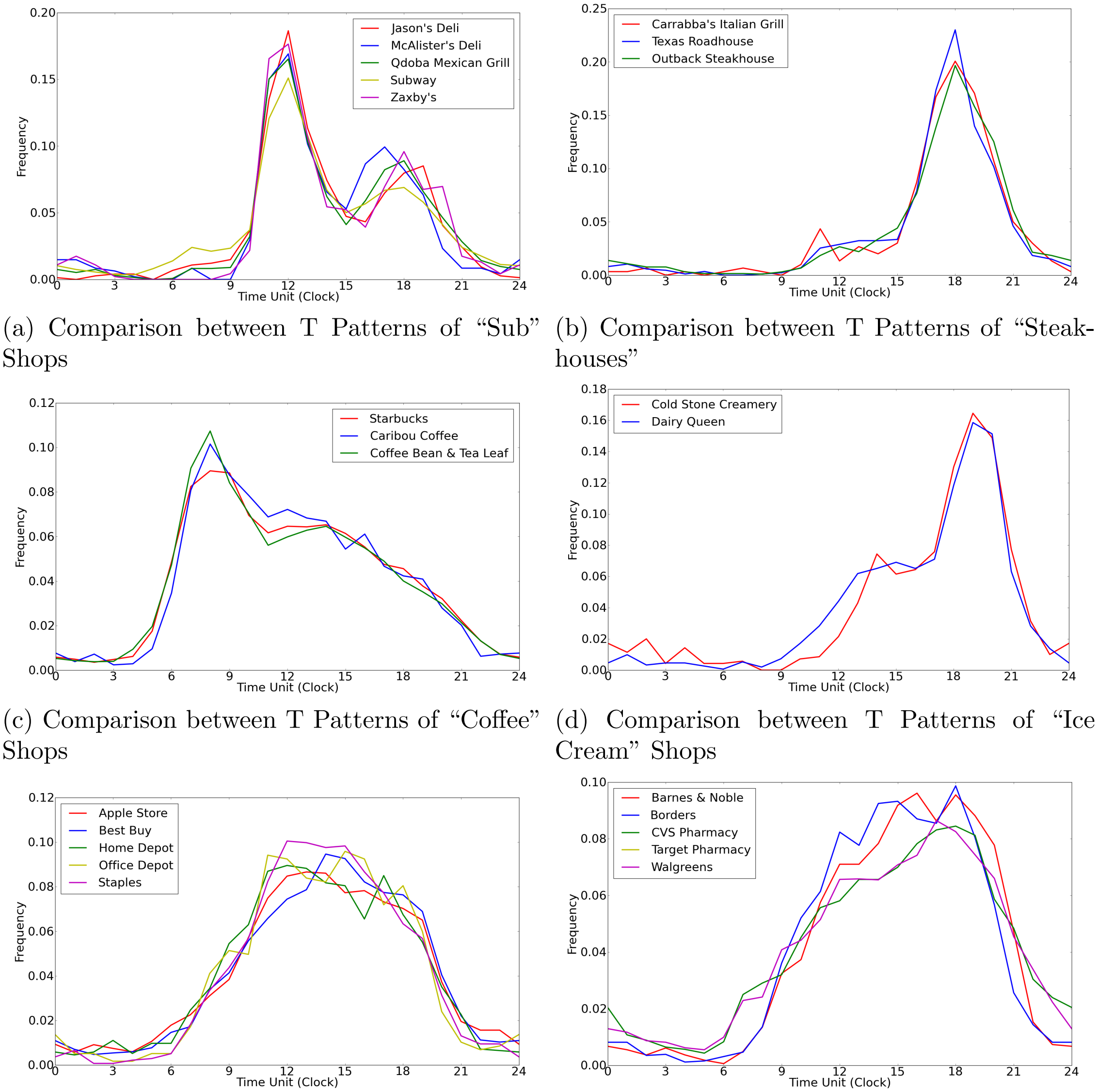
As an example, when we set the pre-defined threshold for minimum temporal similarity as 0.93, a graph with 68 vertices, and 12 strong connected components is generated. Six example components are plotted in Figure 5.4. One component (plotted in Figure 5.4a) contains "Jason's Deli", "McAlister's Deli", "Qdoba Mexican Grill", "Subway", and "Zaxby's" which are all chain restaurants. The traffic patterns of the "steakhouse" component is displayed in Figure 5.4b. Both the sub restaurants and the steakhouses have two peaks (lunch time and dinner time), though the frequencies differ dramatically. The major crowd arrives at sandwich shops at noon to grab lunch, while many more people choose to have steaks for dinner rather than lunch. Similar comparisons are plotted in Figure 5.4c and Figure 5.4d, in which traffic patterns for the component of coffee shops and ice cream shops are featured respectively. People buy coffee in the early morning, and the busyness for the coffee shops gradually decreases during the day. However, people tend to have ice cream in the afternoon, and the ice cream shops are especially crowded in the early evening after dinner. Figure 5.4e and Figure 5.4f plot the components corresponding to office supply stores, and to a component of book store and pharmacies. Both the components have quite similar traffic patterns, and traffic patterns for book stores and pharmacies decrease from their peaks slower than the traffic patterns for the group of office supply stores. The other strong connected components are also highly semantically connected: there is a group of pizza restaurants, a group of juice shops, a group of chained family restaurants, and two groups of fast food restaurants. All the examples validate our hypothesis that measuring temporal correlation between traffic patterns can identify groups of venues that are semantically connected.

5.5 Clustering Venues

Motivated by the results shown in last section, a natural step to utilize the traffic patterns is to group the venues into similar categories. For example, clustering the traffic patterns may lead to clusters of categorized groups of venues, such as “coffee shops”, “steakhouses”, “hotels”, and “gyms”. Specifically, in this section, we apply several different clustering methods and different similarity metrics to cluster the venues based on features of traffic patterns.

5.5.1 Methods

The graph modeling method we used in mining the semantically group of venues is one way to cluster the venues. However, it heavily relies on the value of the



(a) Comparison between T Patterns of "Sub" Shops

(b) Comparison between T Patterns of "Steakhouses"

(c) Comparison between T Patterns of "Coffee" Shops

(d) Comparison between T Patterns of "Ice Cream" Shops

(e) Comparison between T Patterns of "Office Supply Stores"

(f) Comparison between T Patterns of "Book Stores and Pharmacies"

Figure 5.4: Comparisons between T Patterns in Different Strong Connected Groups

pre-defined correlation threshold, and it is difficult to control the number of strong connected components by tuning the threshold. Thus, in this section, we apply K-means and EM clustering algorithm for grouping the venues given their traffic patterns. We apply features of daily traffic pattern, weekly traffic pattern, and daily

+ weekly traffic pattern respectively in clustering the venues. We also explore using the tags for the venues to group the venues into semantically correlated categories.

5.5.2 Experimental Setup

To evaluate the clustering results, a ground truth category is retrieved for each of the venues from the venues dataset collected from Foursquare. For K-means algorithm, we apply Euclidean Distance, Manhattan Distance, and Temporal Distance. Worth mentioning, the temporal correlation discussed earlier ranges between -1 and 1, with higher values indicating higher correlation. Since clustering requires a distance measure, we rescale the temporal correlation to define a metric **Temporal Distance**, where $T_{Dist} = 1.0 - \frac{T_{Corr}+1}{2}$. This temporal distance ranges between 0 and 1, with larger values indicating less similarity (greater distance). In evaluating the results for the clustering, we use the F1-Measure and Purity, which are both standard metrics to evaluate the quality of a clustering. Specifically, F1-Measure balances both the precision and recall of clustering. Purity measures the ratio of the total number of correctly clustered venues over the total number of venues.

Additionally, four test sets are generated to evaluate the clusters based on the criteria of minimum number of check-ins (500+, 300+, 200+, and 100+) for each venue. Venues with more check-ins are expected to have denser traffic patterns, and thus contain stronger indication of semantic information. The four test sets include 148, 242, 383, 585 venues respectively. For each venue, two feature sets are generated: traffic patterns, and vector space models generated from tags. For the traffic patterns, we use daily traffic pattern, weekly traffic pattern, and daily plus weekly traffic pattern respectively. For the vector space models, we retrieve tags for the venues from the venue dataset, and generate features of tf, idf, and tf-idf values for the tags respectively. To normalize the features, we apply L2 normalization on both vectors of traffic patterns and for the vector space model. In the four testsets, only venues of the categories "Food", "Shop", "Home, Work, and Other", and "Travel Spot" (among the eight categories in categorization system of Foursquare) have sufficient check-ins. Thus, in both K-means and EM algorithm, we pre-specify the number of clusters to be 4.

5.5.3 Experimental Results

5.5.3.1 Clustering with the Traffic Patterns

Results for K-means and EM algorithms evaluated by F1-Measure and Purity using the four test sets (with 500+, 300+, 200+, and 100+ check-ins respectively) are listed in Table 5.3. In most of the cases, clustering with the daily traffic pattern itself performs the best. Combining both daily and weekly traffic patterns performs better than using weekly traffic pattern alone. Among combination of method and metric, K-means plus Temporal Distance performs the best with a significant increase (around 15% to 20% increase) over K-means with the other two metrics, as well as the EM algorithm. K-means with Manhattan Distance performs better than K-means with Euclidean Distance. And they both outperform the EM algorithm when the venues have the sufficient check-ins. However EM algorithms performs competitively when the venues have fewer check-ins.

The best performing methods and their results are extracted from Table 5.3 and plotted in Figure 5.5. As we can see from Figure 5.5, generally, test sets with denser traffic patterns reach better performance in F1-Measure, and Purity. Results for the dataset with least check-ins suffer from lack of sufficient data in traffic patterns. We also observe that the dataset with the most check-ins do not reach the best performance. This is partly due to the lack of number of venues (only 2) in the category of travel spots.

Table 5.3: Results for Traffic Pattern Clustering

Dataset	Features	Daily T Pattern		Weekly T Pattern		Daily + Weekly T Pattern	
	Method / Metric	F1	Purity	F1	Purity	F1	Purity
500+	K-means + Euclidean	0.495	0.595	0.419	0.534	0.412	0.520
	K-means + Manhattan	0.502	0.608	0.442	0.554	0.451	0.574
	K-means + T Distance	0.539	0.608	0.603	**0.689**	**0.608**	0.635
	EM	0.424	0.541	0.43	0.554	0.416	0.534
300+	K-means + Euclidean	0.427	0.504	0.465	0.533	0.466	0.562
	K-means + Manhattan	0.502	0.566	0.435	0.533	0.476	0.583
	K-means + T Distance	**0.675**	**0.764**	0.586	0.674	0.616	0.698
	EM	0.465	0.537	0.467	0.579	0.47	0.562
200+	K-means + Euclidean	0.416	0.483	0.463	0.441	0.456	0.535
	K-means + Manhattan	0.498	0.527	0.446	0.504	0.477	0.548
	K-means + T Distance	**0.671**	**0.700**	0.566	0.621	0.527	0.585
	EM	0.482	0.512	0.412	0.452	0.408	0.446
100	K-means + Euclidean	0.415	0.525	0.435	0.535	0.427	0.444
	K-means + Manhattan	0.437	0.535	0.406	0.504	0.437	0.515
	K-means + T Distance	0.571	0.667	0.552	0.658	**0.599**	**0.706**
	EM	0.477	0.568	0.403	0.487	0.464	0.405

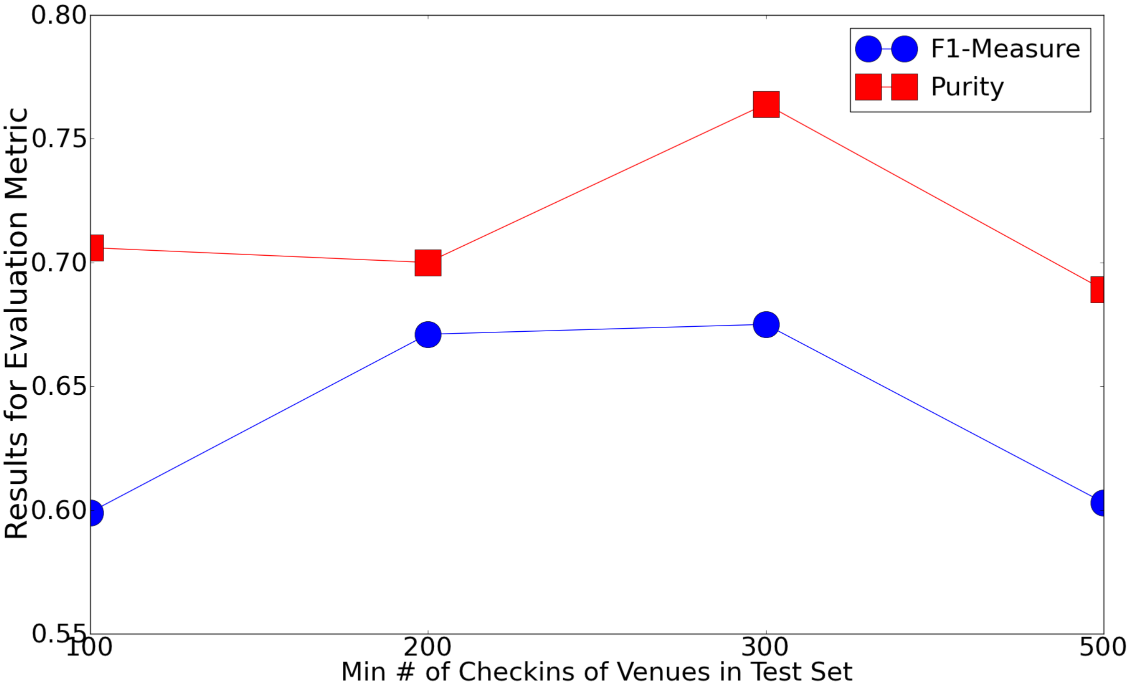


Figure 5.5: Comparison of Best Results over Different Test Sets

5.5.3.2 Clustering with the Vector Space Models

As mentioned, we retrieve tags for the venues in the four test sets from the venue dataset collected from Foursquare. We apply the same clustering methods using the features of vector space models (tf, idf, tf-idf) modeled from the tags respectively. However, for different methods over different training sets, all the venues tend to converge to a single large cluster, primarily due to the sparseness of the tags in the datasets.

Together, these results show that modeling venues by traffic patterns and using temporal correlation to measure venue similarity are viable alternatives to traditional content-based clustering methods.

5.6 Supervised Venue Categorization

While clustering of venues by traffic-based temporal correlation can provide a foundation for organizing venues for improved location-based search, there is still the challenge of sparsity for venues on the "long tail". Hence, in this section, we propose and evaluate a traffic pattern-driven approach for supervised location categorization, wherein traffic patterns can be used to accurately predict the semantic category of uncategorized locations. Given a set of venues with known category labels, and their corresponding traffic patterns, are we able to train classifiers with the labeled set to categorize incoming venues with their corresponding traffic patterns? As in our study of clustering, we also consider an alternative tag-based model as a point-of-comparison.

5.6.1 Training Set and 10-Fold Cross Validation

The training set contains a set of the 271 most popular venues which all have at least 100 branches and over 100 check-ins in the check-ins dataset. As mentioned earlier, we retrieve the ground-truth labels for the venues from Foursquare venue data. The 271 venues are grouped into four categories (all belong to the category system mentioned before): Food; Home, Work & Other; Shop; and Travel Spot. We adopt the same two set of features as in the clustering work: traffic pattern (including daily traffic pattern weekly traffic pattern, and daily plus weekly traffic patterns), and vector space models (tf, idf, and tf-idf values) for tags associated to the venues.

We apply the features in training classifiers of Naive Bayes, 1NN (we iterate k from 1 to 5 for kNN, and 1NN always perform the best, so we fix k to 1 in the following experiments), AdaBoostM1, and SimpleCart. We also apply L2 normalization on both vectors of traffic patterns and for the vector space model. To evaluate the effectiveness of the classifiers, we use 10-fold cross validation and the F1-Measure.

5.6.1.1 Traffic Pattern Features

Results for traffic pattern features are plotted in Figure 5.6. Using either daily traffic pattern or the weekly traffic pattern displays a strong indication of the category of the venue, reaching an F1-Measure higher than 0.8 with 1NN classifier. The other three classifiers – Naive Bayes, AdaBoostM1, and Simple Cart – seem incompatible with the time series data. Combining both daily traffic pattern and weekly traffic pattern gives a boost of 6.5% for Simple Cart, 1.7% for 1NN, and 0.6% for Naive Bayes. Overall, the 1NN classifier still performs best with an F1-Measure of 0.820 slightly above AdaBoostM1 and SimpleCart with a best F1-Measure of 0.819. These results agree with the observations in Xi et al.'s work [121] that 1NN classifier has excellent performance in time-series classification.

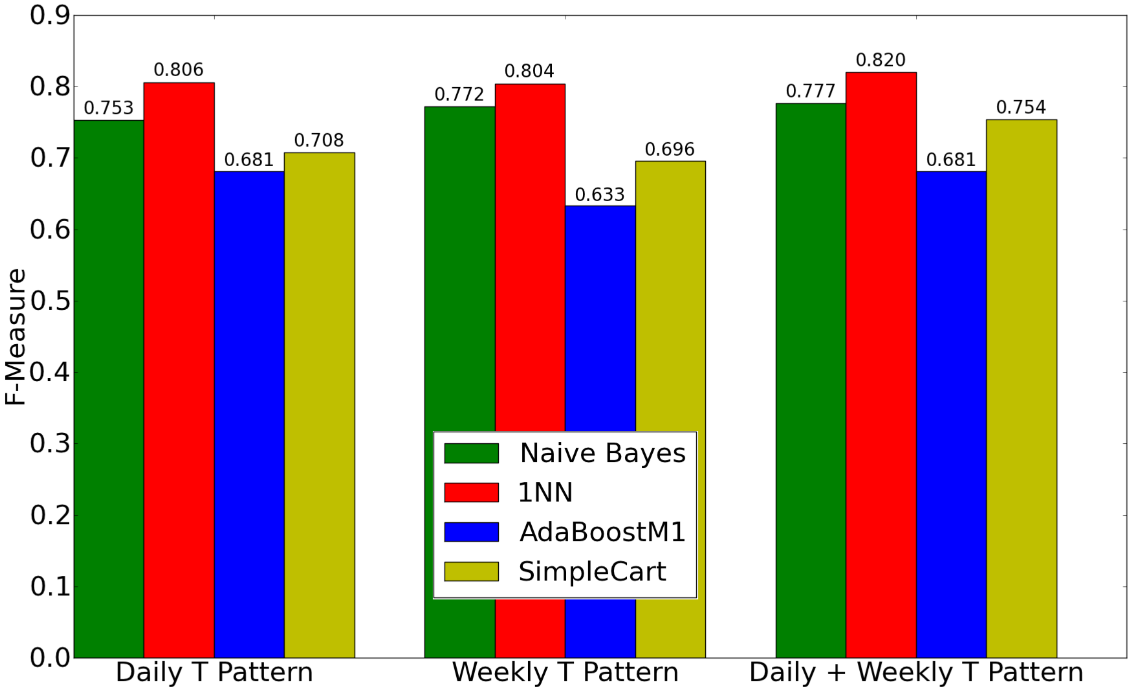


Figure 5.6: 10-Fold Cross Validation of 271 Venues Classification With T Patterns

5.6.1.2 Vector Space Model Features

Traditionally, an alternative way to categorize locations is using the social tags associated with the locations. As in the clustering approach, we again apply the vector space model features using tf, idf, and tf-idf respectively. Results in Figure 5.7 significantly differ from the results shown in Figure 5.6: for both tf and tf-idf, Naive Bayes, AdaBoostM1, and SimpleCart perform much better than 1NN, which verifies our previous assumption that those classifiers work well with text-driven features. Classifiers trained by tf or tf-idf models significantly outperform ones trained by idf models. Furthermore, classifiers trained by tf-idf and tf have very similar results, which shows that enriching with the idf information could not help classify the venues. Besides, the best result for the vector space model based classifiers performs a little better (1.3%) than the best result for traffic pattern based classifiers.

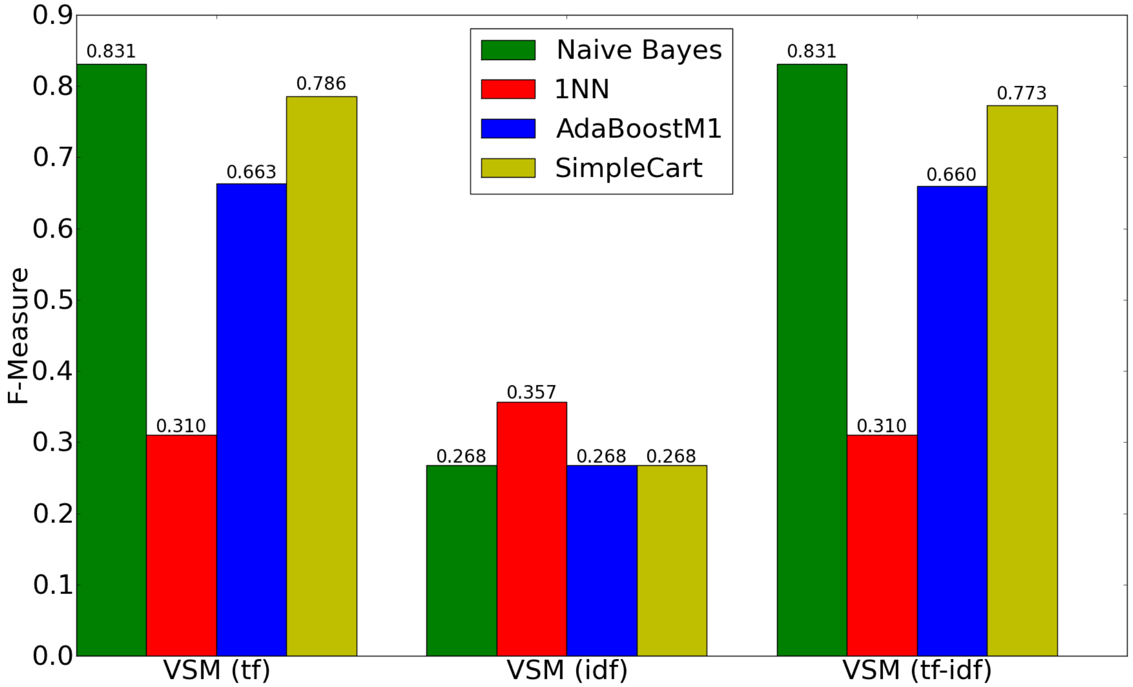


Figure 5.7: 10-Fold Cross Validation of 271 Venues Classification With Vector Space Models

5.6.1.3 Combination of the Two Sets of Features

To answer the question of whether enriching information from tags could help facilitate venue categorization, we train classifiers on both sets of features (each set of features are normalized independently with L2-normalization before being merged). Unexpectedly, the results using both sets of features (daily + weekly t pattern, and tf-idf vector models) perform similar or even a little bit worse than the best performing classifiers trained by either set of features (results listed in Table 5.4). We attribute this drop in performance to the inclusion of possibly unhelpful countervailing features (94 temporal features, and 10,324 vector space model features). Thus, we apply feature selection to reduce the number of features by filtering irrelevant and redundant features.

Table 5.4: 10-Fold Cross Validation of 271 Venues Classification With Traffic Pattern and Vector Space Models

Metric	Naive Bayes	1NN	AdaBoostM1	SimpleCart
F1-Measure	0.831	0.59	0.66	0.753

5.6.1.4 Feature Selection

We apply the standard Chi-Square feature selection method to reduce the number of vector space model features from 10,324 to 358 by filtering insignificant and redundant features. All traffic pattern features are considered significant by Chi-Square and so remain for the following classification experiments. Thus, we get exactly same results for traffic patterns comparing to results in Figure 5.6. The results for vector space model features are displayed in Figure 5.8. Comparing to the results without feature selection, classifiers trained by idf features, AdaboostM1 trained by tf features, and tf-idf features still perform the same. 1NN classifier trained with tf features and tf-idf features outperform with an almost 158.4% increase over the previous results. Besides, Naive Bayes classifiers trained with tf features and tf-idf features also have a 3.2% increase in their performance; as well as the Simple Cart with over 1% increase. The best performing classifier so far becomes Naive Bayes, which reaches a F1-Measure of 0.861.

Feature selection also shows its effectiveness when we train the classifiers using both the traffic pattern features and vector space model features. Comparing to results in Table 5.4, results in Table 5.5 show increase of performance for 1NN (29.7%), and Naive Bayes (4.2%).

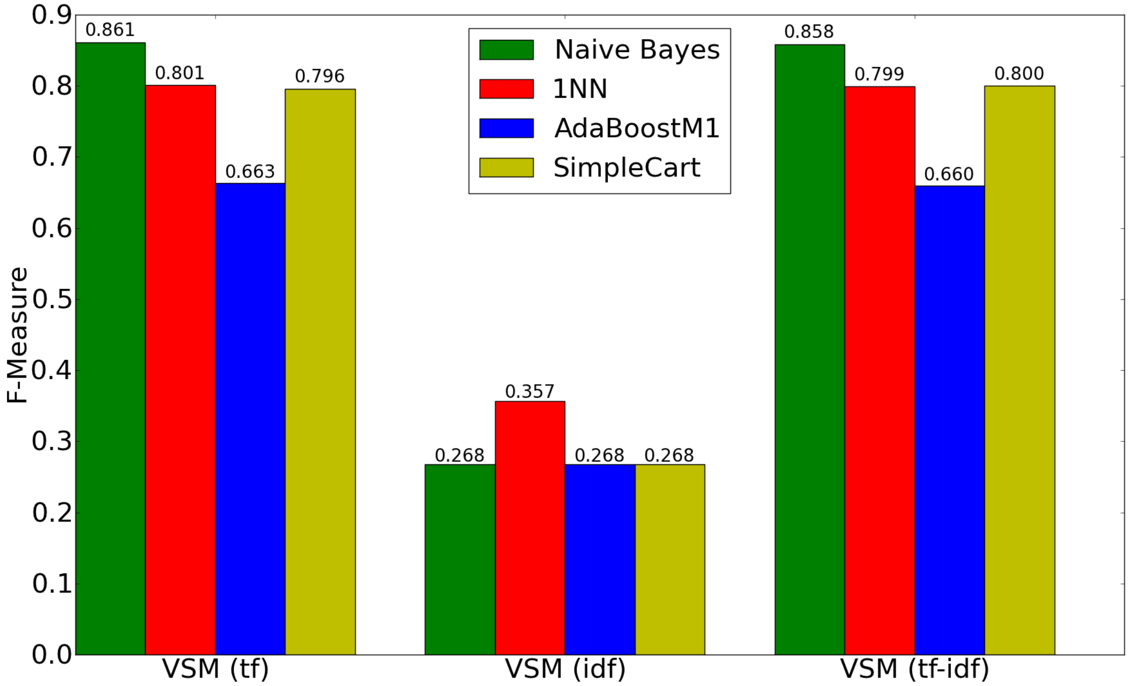


Figure 5.8: 10-Fold Cross Validation of 271 Venues Classification With Vector Space Models + Feature Selection

Table 5.5: 10-Fold Cross Validation of 271 Venues Classification With Traffic Pattern and Vector Space Model + Feature Selection

Metric	Naive Bayes	1NN	AdaBoostM1	SimpleCart
F1-Measure	0.866	0.765	0.66	0.75

5.6.2 Evaluation on Test Set

Based on the 10-fold cross validation on the training set, we only use 1NN as the classifier to classify venues based on the feature of traffic patterns. The test set of venues are generated based on the criteria of at least 10 branches with certain number of check-ins above a pre-defined threshold. We set the threshold as 10, 30, 50, 100, 200, 300, and 500 respectively, and the corresponding number of venues in the test sets are listed in Table 5.6. As we can see from the table, with a relaxed criteria of only 10 check-ins and above, the test set contains 1,392 venues, and with a strict criteria of 500 check-ins and above, the test set only contains 21 venues. Note that the test sets are disjoint with the training set.

We train the 1NN classifier on the training set using daily traffic pattern, weekly traffic pattern, and daily plus weekly traffic patterns respectively, and evaluate test sets with corresponding features. The results for the classification are plotted in Figure 5.9 (each tick in x axis represents a test set). As we can see in the figure, with more check-ins required for a venue in a test set, the results get better for the classifier trained by daily traffic pattern, and it reaches its peak with an F1-Measure of 0.742. However, it gets worse performance for the tests requiring at least 300 check-ins and 500 check-ins. This is probably caused by the lack of venues in the two test sets, and a small number of mis-classified venues can significantly affect the results. For the classifier trained by weekly traffic pattern and daily plus weekly traffic pattern, the results generally get better with test sets which only contain venues with dense traffic patterns. With weekly traffic pattern features, the classifier works much better overall than the one trained by daily traffic patterns. The classifier trained by daily plus weekly traffic patterns works slightly better than the one trained by weekly traffic patterns with test sets with relaxed condition, falls behind a little bit for test set 100, and test set 200, and finally catches up for test set 300 and test 500. In the figure, we can see that with only 50 or more check-ins input per venue, the 1NN classifier can reach a F1-Measure almost 0.75, which shows its good performance in venue categorization.

Table 5.6: Number of Venues in Test Sets

Min # of Check-ins	10	30	50	100	200	300	500
# of Venues	1392	983	695	353	142	60	21

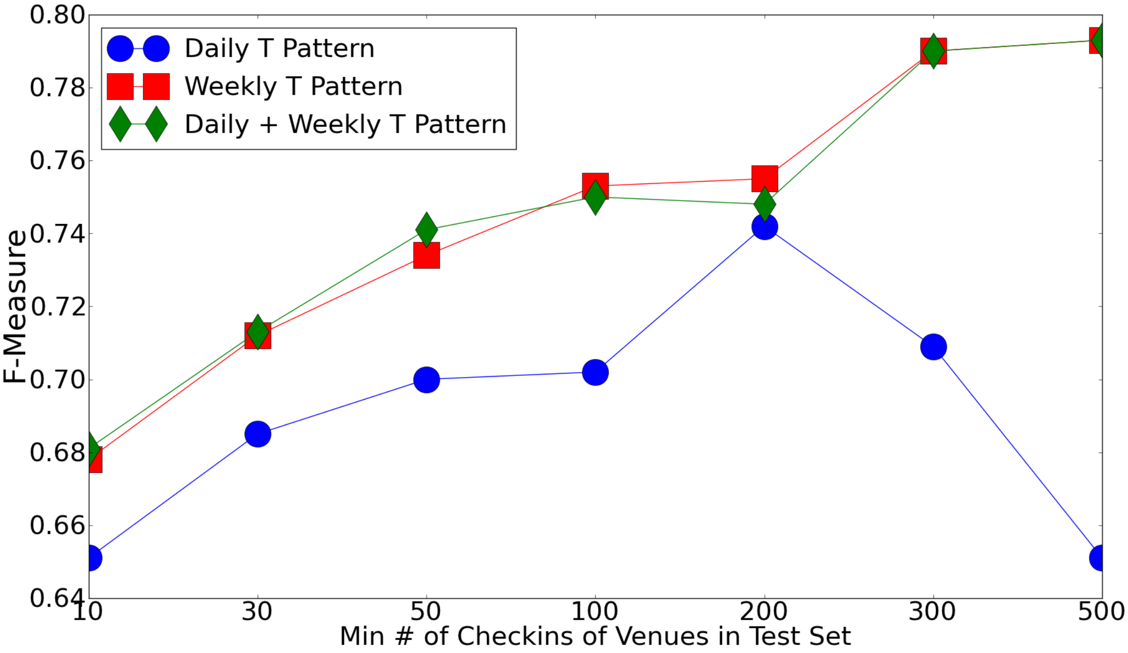


Figure 5.9: Evaluation of 1NN Trained by Traffic Patterns on Test Sets

5.7 Augmenting Location-Based Search

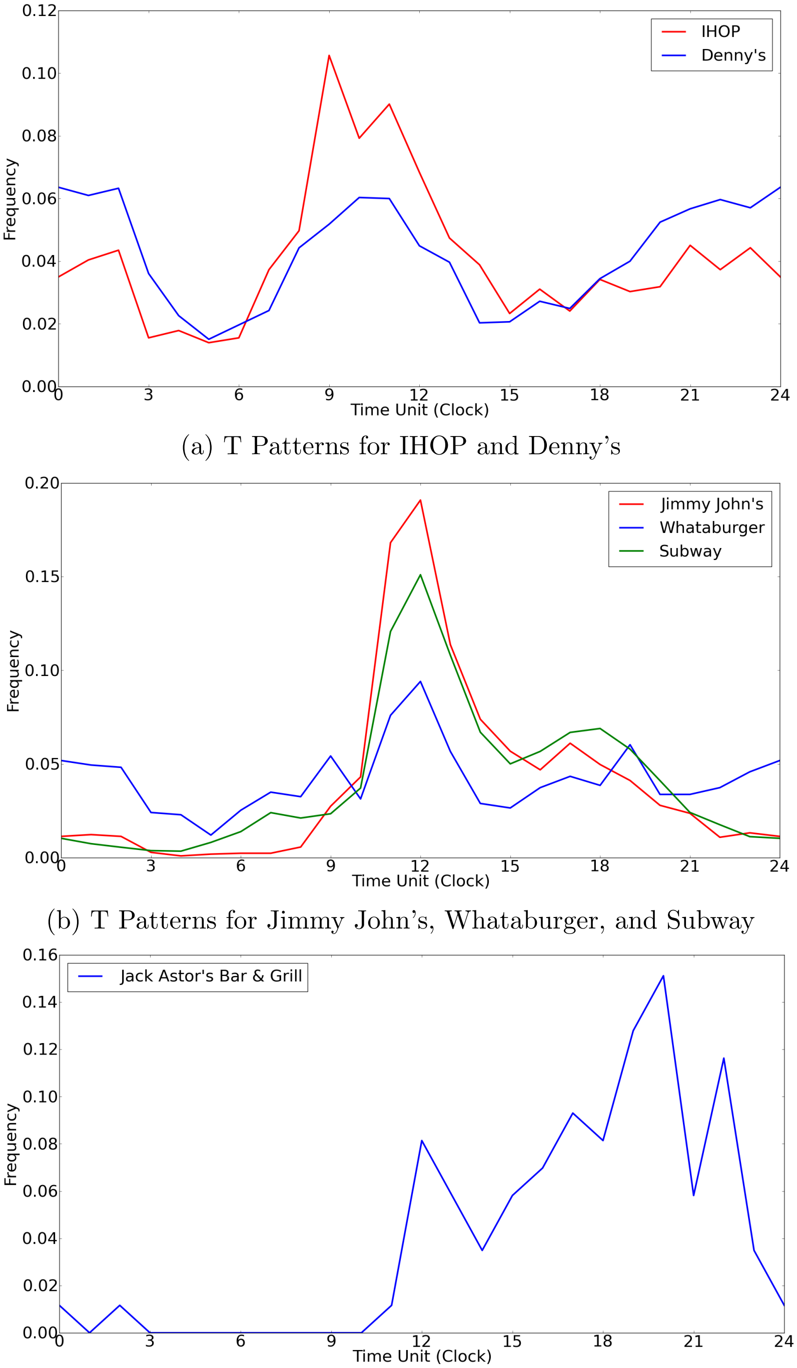
So far, we have seen that the traffic patterns for venues revealed through location sharing services contain rich information about the venues' semantic category. And we have successfully taken advantage of these traffic patterns for both unsupervised semantic group clustering and supervised venue categorization. In this section, we show how we can incorporate venues' traffic patterns and their category information into traditional location-based web search.

5.7.1 Answering Queries for Traffic

Traffic patterns and category information for venues can be easily incorporated into traditional location-based search to answer the information need for traffic. One scenario for answering the traffic-driven query is: Karen is searching for a restaurant which is off-peak during dinner time between 5 - 7 PM, so that she can enjoy the quiet environment talking with her friends. Knowing the traffic patterns and category for venues, the system could easily retrieve the venues nearby in the category of food, and rank the results by the descending order of busyness. Example results are plotted in Figure 5.10. For example, Karen can choose IHOP and Denny's where the crowd usually come in the early morning, lunch time, and late in the evening; she can also go to fast food venues like Jimmy Johns, and Chipotle which are crowded in during lunch time; besides, Karen can also choose grill & bars which are more popular in late evening like Jack Astor's Bar & Grill.

5.7.2 Location Recommendation based on Traffic

Another potential application is recommendation of venues having similar traffic patterns. For example, Jerry plays a lot of basketball, and tennis. He usually goes to the Williams Park during Wednesday early evening, and Saturday afternoon, which are both free time for him and peak times for guys to get-together and play basketball and tennis. Recently, he moves to a new neighborhood, and wants to find places nearby that have similar traffic patterns, so that he can meet new friends there and play some basketball or tennis. A traffic-driven location-based search can also easily handle this kind of queries. Given the name of the venue, the system calculates temporal similarity between traffic patterns of the venue and other venues in the same category (or in other categories as well), and return the locations with the highest temporal similarities. The example results are plotted in Figure 5.11,



(a) T Patterns for IHOP and Denny's

(b) T Patterns for Jimmy John's, Whataburger, and Subway

(c) T Pattern for Jack Astor's Bar & Grill

Figure 5.10: Traffic Patterns for Venues Off-Peak between 5-7 PM

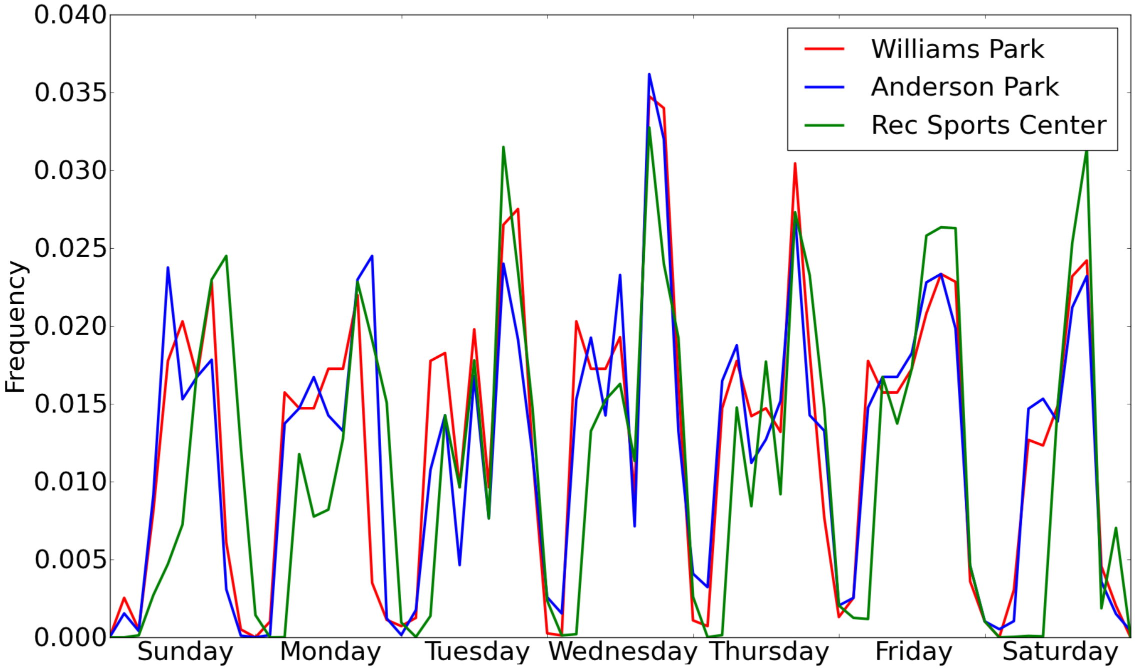


Figure 5.11: Example Showing Venue Recommendation based On T Pattern

which shows the comparison of traffic patterns of Williams Park and two similar nearby venues Anderson Park and Rec Sports Center.

5.8 Summary

In this section, we propose to mine activity patterns revealed through location sharing services to augment traditional location-based search. Strong indication of semantic information are found in the activity patterns generated from 22 million check-ins from location sharing services. Then, we take advantage of the activity patterns and successfully cluster venues into semantically correlated groups, and categorize incoming venues based on the associated activity dynamics. Based on the results, we also provide two examples to show how activity-driven semantic organization of locations may be naturally incorporated into traditional location-based search.

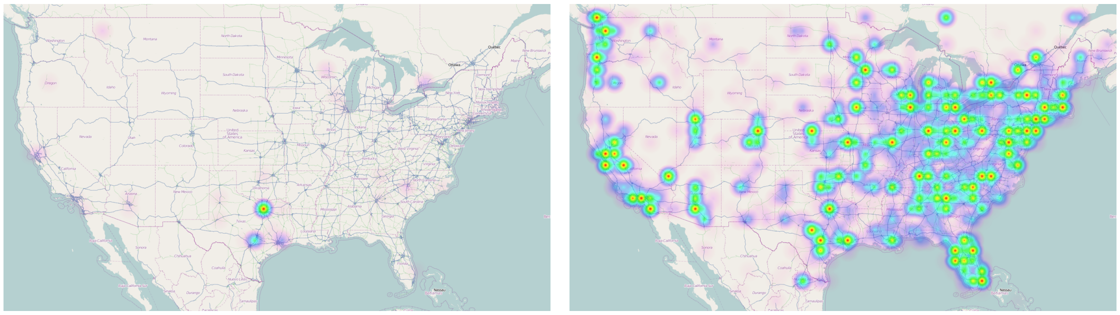
6. A GEO-SPATIAL APPROACH TO FINDING LOCAL EXPERTS ON TWITTER

6.1 Introduction

We tackle the problem of finding *local experts* in social media systems like Twitter. Local experts bring specialized knowledge about a particular location and can provide insights that are typically unavailable to more general topic experts. For example, a "foodie" local expert is someone who is knowledgeable about the local food scene, and may be able to answer local information needs like: what's the best barbecue in town? Which restaurants locally source their vegetables? Which pubs are good for hearing new bands? Similarly, a local "techie" expert could be a conduit to connecting with local entrepreneurs, identifying tech-oriented neighborhood hangouts, and recommending local talent (e.g., do you know any good, available web developers?). Indeed, a recent Yahoo! Research survey found that 43% of participants would like to directly contact local experts for advice and recommendations (in the context of online review systems like Yelp), while 39% would not mind being contacted by others [3].

And yet finding local experts is challenging. Traditional expert finding has focused on either small-scale, difficult-to-scale curation of experts (e.g., a magazine's list of the "Top 100 Lawyers in Houston") or on automated methods that can mine large-scale information sharing platforms. Indeed, many efforts have focused on finding experts in online forums [133], question-answering sites [76], enterprise corpora [8, 16], and online social networks [19, 46, 90, 119, 132]. These approaches, however, have typically focused on finding general topic experts, rather than *local experts*.

In this section, we investigate new approaches for mining local expertise from social media systems like Twitter. Our approach is motivated by the widespread adoption of GPS-enabled tagging of social media content via smartphones and social media services (e.g., Facebook, Twitter, Foursquare). These services provide a *geo-social* overlay of the physical environment of the planet with billions of check-ins, images, Tweets, and other location-sensitive markers. This massive scale geo-social resource provides unprecedented opportunities to study the connection between people's expertise and locations and for building localized expert finding systems.



(a) @BBQsnob

(b) @JimmyFallon

Figure 6.1: Heatmap of the Location of Twitter Users Who Have Listed @BBQsnob or @JimmyFallon

Concretely, we propose a local expertise framework – LocalRank – that integrates both a person's topical expertise and their local authority. The framework views a local expert as *someone who is well recognized by the local community*, where we estimate this local recognition via a novel spatial proximity expertise approach that leverages over 15 million geo-tagged Twitter lists. To illustrate, Figure 6.1a shows a heatmap of the locations of Twitter users who have labeled Daniel Vaughn (@BBQsnob) on Twitter. Vaughn – the newly-named Barbecue Editor of Texas Monthly – is one of the foremost barbecue experts in Texas. We can see that his expertise is recognized regionally in Texas, and more specifically by local barbecue centers in Austin and Dallas. In contrast, late-night host Jimmy Fallon's heatmap suggests he is recognized nationally, but without a strong local community. Intuitively, Daniel Vaughn is recognized as a *local expert* in Austin in the area of Barbecue; Jimmy Fallon is certainly an expert (of comedy and entertainment), but his expertise is diffused nationally.

Toward identifying local experts, this section makes the following contributions.

• First, we propose the problem of *local expert finding* in social media systems like Twitter and propose a novel expertise framework – LocalRank. The framework decomposes local expertise into two key components: a candidate's topical authority (e.g., how well is the candidate recognized in the area of Barbecue or web development?) and his local authority (e.g., how well do people in Austin – the area of interest – recognize this candidate?).

• Second, to estimate *local authority*, we mine the fine-grained geo-tagged linkages among millions of Twitter users. Concretely, we extract Twitter list relationships where both the list creator and the user being labeled have revealed a precise location. The first local authority method considers the distance between an expert candidate's location and the location of interest, capturing the intuition that closer candidates are more locally authoritative. However, in many cases, an expert in one location may actually live far away – e.g., Daniel Vaughn is an expert in Austin Barbecue although he lives 200 miles away in Dallas. To capture these cases, we propose and evaluate a local authority method that considers the distance of the candidate expert's "core audience" from the location of interest (that is, to reward candidates who have many labelers near the location of interest, even if the candidate lives far away). So, if many people in Austin consider Daniel Vaughn an expert, then his Austin local authority should reflect that.

- Third, to estimate *topical authority*, we adapt a well-known language modeling approach to expertise identification, but augment it to incorporate the distance-weighted social ties of 24 million geo-tagged Twitter users. In this way, topical expertise can be propagated through the social network to identify local experts that are well connected to, and recognized by the local community in the topic.

- Finally, we evaluate the LocalRank framework across 56 local expertise queries coupled with 11,000 individual judgments from Amazon Mechanical Turk. We see a significant improvement in performance (35% improvement in $Precision@10$ and around 18% in $NDCG@10$) over the best performing alternative approach. We observe that the local authority approaches that consider the locations of a candidate's "core audience" perform much better than an alternative that only considers the distance between the candidate's location to the query location. In addition, we see that the expertise propagation through the social network can improve the baseline local expert finding approach.

These results demonstrate the viability of mining fine-grained geo-social signals for expertise finding, and highlight the potential of future geo-social systems that facilitate information flow between local experts and the local community.

6.2 Related Work

The emergence of online geo-social systems provides unprecedented opportunities to bridge the gap between people's online and offline presence [22, 59]. However there are key challenges associated with these opportunities including location sparsity [7, 17, 78] and location privacy [26, 130]. Given the geo-social footprints from these services, researchers have analyzed the spatio-temporal properties of these footprints [79, 103, 104], studied the semantics associated with these footprints [127, 94], and investigated new location recommendation systems [87, 129, 134, 135].

Expert finding is an important task that has seen considerable research. Lappas et al. [66] provided a comprehensive survey about expert finding in social networks, and grouped the related work into two categories: (i) using text content posted by expert candidates; and (ii) using the expert candidates' online social connections. For example, Balog et al. [8] proposed two generative probabilistic models – a user model generated using documents associated to an expert, and a topic model generated using documents associated to the topic – to detect topic experts. Based on their evaluation over the TREC Enterprise corpora, the authors observed that the topic model outperforms the user model and other unsupervised techniques. On the other hand, Zhang et al. [133] applied link analysis approaches like PageRank and HITS to identify top experts in a Java forum, observing that both link analysis and network structure are helpful in finding users with extensive expertise.

Along the direction of expert finding in online social networks, Weng et al. [119], proposed a link-analysis based approach to identify top experts in a topic. They considered both topical similarity between users and social connections. The authors observed their approach outperforms Twitter's system, PageRank, and topic-sensitive Pagerank. Similarly, Pal and Counts [90] introduced a probabilistic clustering framework to identify top authorities in a topic using both nodal and topical features. The Cognos system built by Ghosh et al. [46] leveraged Twitter lists to identify the candidate's expertise, and the authors reported that their system works as well as Twitter's official system (i.e., WTF: Who To Follow) to identify top users for a particular topic. Other works include expert finding in online forums [133], question-answering sites [76], enterprise corpora [16, 8], and online social network services [19, 46, 90, 119, 132].

In the context of local experts, Antin et al. [3] recently presented a survey designed to examine people's attitudes about local knowledge and personal investment in local neighborhoods. They observed that over 52% of the participants claimed having both local knowledge and personal investment in their local area. And in an encouraging direction, they found that many participants would like to contact local experts for advice (43%) and many would not mind being contacted by others (39%).

6.3 LocalRank: Problem Statement and Solution Approach

In this section, we are interested to find local experts with particular expertise in a specific location. We assume there is a pool of expert candidates $V = \{v_1, v_2, ..., v_n\}$, that each candidate v_i has an associated location $l(v_i)$ and a set of areas of expertise described by a feature vector $\vec{v_i}$. Each element in the vector is associated with a expertise topic word t_w (e.g., "technology"), and the element value indicates to what extent the candidate is an expert in the corresponding topic. We define the **Local Expert Finding** problem as:

Definition 1 ***(Local Expert Finding)*** *Given a query q that includes a query topic $t(q)$, and a query location $l(q)$, find the set of k candidates with the highest local expertise in query topic $t(q)$ and location $l(q)$.*

A location $l(q)$ can correspond to different spatial granularities, depending on the goal of expert finding – a region (e.g., Texas), a city (e.g., Austin), a neighborhood (e.g., downtown), or a latitude-longitude coordinate.

6.3.1 Topical vs. Local Authority

Identify a local expert requires that we can accurately estimate not only the candidate's expertise on a topic of interest (e.g., how much does this candidate know about barbecue), but also that we can identify the candidate's local authority (e.g., how well does the local community recognize this candidate's expertise). Hence,

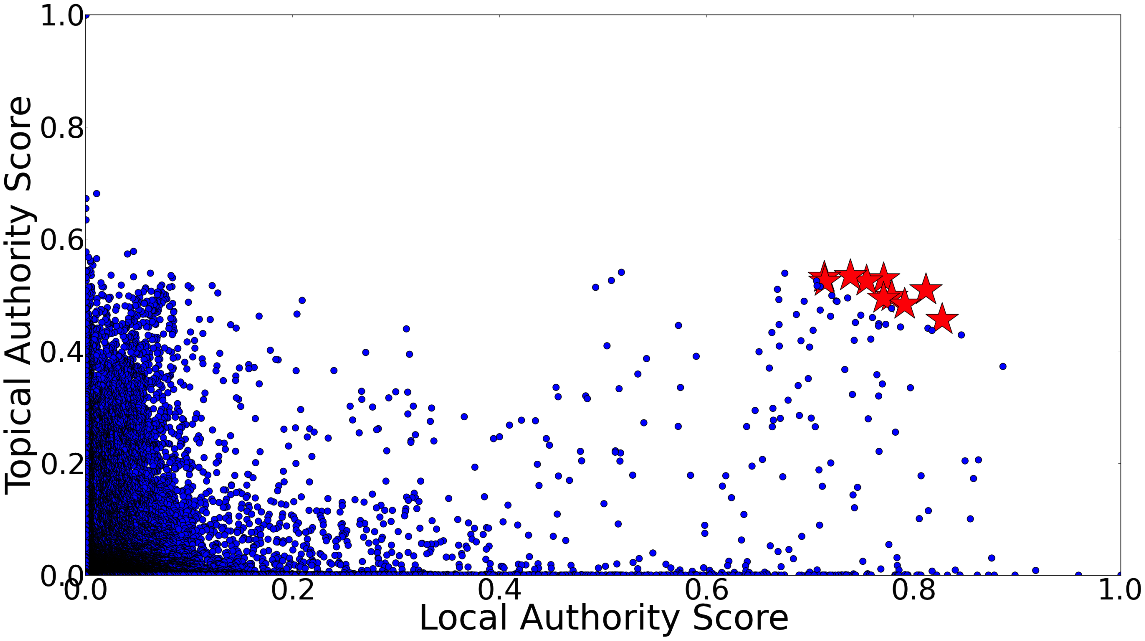


Figure 6.2: Our Goal is to Identify Local Experts (the Red Stars in the Top-right Section)

we propose to decompose the local expertise for a candidate v_i into two related dimensions:

- **Topical Authority**: which captures the candidate's expertise on the topic area $t(q)$.
- **Local Authority**: which captures the candidate's local authority in query location $l(q)$.

To illustrate, Figure 6.2 shows example candidates in this two-dimensional space for a particular topic (say, Barbecue) and a particular location (say, Austin):

- *Nobodies [bottom-left]*: For a particular area of interest, these candidates have both low topical authority and low local authority.

- *Locals [bottom-right]*: These candidates have high local authority, but low topical authority. E.g., an author or artist living in Austin.
- *Experts [top-left]*: Candidates with high topical authority, but low local authority. These candidates are certainly experts on a topic, but are not well recognized locally for this expertise. E.g., an expert in pork barbecue originating in North Carolina, but not beef barbecue in Texas.
- *Local Experts [top-right: red stars]*: both great topical authority and local authority. E.g., Daniel Vaughn, the Barbecue Editor of Texas Monthly.

Note that a candidate is evaluated per-topic and per-location, so a local expert in one place may be considered as just an expert or even a nobody in a different location.

6.3.2 Local Expertise in Twitter Lists

To identify these local experts (the red stars), we propose to exploit the geo-social information embedded in Twitter lists to find candidates who are *well recognized by the local community.* Twitter lists are a form of crowd-sourced knowledge, whereby aggregating the individual lists constructed by distinct users can reveal the crowd perspective on how a Twitter user is perceived [46]. Concretely, for each expert candidate v_i, we assume that there is a set of people $V_l(v_i)$ that recognize v_i's expertise, and label v_i in their own lists. We refer to the set of people as candidate v_i's *list labelers* or more concisely *labelers.* Candidate v_i is the *labelee.* Critical for our study, for each labeler v_j (such that $v_j \in V_l(v_i)$), we assume that v_j's location $l(v_j)$ is known.

But how do we sample such geo-tagged list relationships? Are there sufficient users to support local expertise finding? And if so, do these lists actually reveal topics of potential expertise interest, or are they focused mainly on other dimensions (e.g., for organizing a user's friends)? In the following, we present our Twitter geo-tagged data collection (summarized in Table 6.1) and address the potential of geo-tagged lists to support local expertise finding, before turning to the development of our local expert finding approach.

6.3.2.1 Geo-Locating Users

We sample 54 million Twitter user profiles based on the ID range of 12 (starting from Twitter co-founder Jack Dorsey @Jack) to 100 million, as well as 3 billion geo-tagged tweets [58]). For each user, we seek to assign a *home location*; however, it is widely observed that many Twitter users reveal overly coarse or no location at all in the self-reported location field (see, e.g., [17]). Hence, we adopt a home finding method that relies on a user's geo-tagged tweets akin to a similar approach previously used for check-ins and geo-tagged images [18, 86]. First, we group the user's locations where he posted his tweets into squares of one degree latitude by one degree longitude (covering about 4,000 square miles). Next, we select the square containing the most geo-tagged tweets as the center, and select the eight neighboring squares to form a lattice. We divide the lattice into squares measuring 0.1 by 0.1 square degrees, and repeat the center and neighbor selection procedures. This process repeats until we arrive at squares of size 0.001 by 0.001 square degrees (covering about 0.004 square miles). Finally, we select the center of the square with the most geo-tagged tweets as the "home" of the user. In total, we geo-locate about 24 million out of the 54 million users (about 45.1%) with fine-grained latitude-longitude coordinates.

6.3.2.2 Geo-Labeled List Relationships

Of the 24 million geo-tagged Twitter users, we sample 13 million lists that these users occur on or that these users have created. In total, the 24 million users occur 86 million times in the 13 million lists. Among these 86 million occurrences of a user

Table 6.1: Geo-tagged Twitter Data

Data Type	Total # of Records
User Profiles	53,743,459
Geo-Tagged User Profiles (45.1%)	24,252,450
Lists	12,882,292
User List Occurrences	85,988,377
Geo-Tagged List Relationships (17.2%)	14,763,767
Friendship Links	166,870,858
List-peer Relationships	430,186,408

in a list, almost 15 million of them are geo-tagged, indicating a direct link from a list creator's location to a list member's location. In addition to this network of list relationships, we additionally collect two additional networks around these users: (i) 167 million friendship links connecting these geo-tagged users; and (ii) 430 million links connecting a pair of geo-tagged users that co-occur in the same list.

6.3.2.3 Expertise Potential of List Names

We parse the list names that are associated with all 14 million geo-tagged list labeling relationships (i.e., links connecting list creator to list member). Table 6.2 shows the most frequent unigrams. We are encouraged to see that 15 of the 21 most frequent unigrams are related to either people's expertise or interests (the others focus on friendship and celebrity); as has been observed by Kwak et al. [65], Twitter serves as a form of news media as well as a social network, so there is good potential for expertise mining.

6.3.2.4 Spatial Patterns of Expertise

What do these geo-tagged lists reveal? For four example topics – "tech", "entertain", "travel", and "food" – we plot in Figure 6.3 the cumulative distribution

Table 6.2: Most Frequent Words in List Names of Geo-tagged List Labeling Relationships

news	2.66%	**media**	1.87%	**music**	1.71%
twibe	1.27%	**tech**	1.11%	people	1.06%
celeb	1.04%	**social**	1.04%	**sport**	1.01%
design	0.84%	**market**	0.81%	**politic**	0.80%
follow	0.70%	celebrity	0.69%	**food**	0.61%
art	0.58%	**business**	0.55%	friend	0.52%
entertain	0.50%	**web**	0.48%	**travel**	0.47%

of frequency of list labeling relationships over distance. That is, how far apart are list labelers from the list labelees? We observe almost 40% of Twitter users who are labelees in a "food"-relevant list are within a hundred miles to the labelers. However, only about 10% to 15% of the labelees in a list of other three topics are within a hundred miles to their labelers. In addition, the average distance between a pair of list labeler and list labelee for "food" is also much smaller than the average distance for other topics. These observations suggest that certain topics are inherently more "local" and that identifying local experts in topics that are inherently more local could be easier than identifying local experts in other topics.

6.3.3 Local Expert Finding with LocalRank

Based on these encouraging observations – (i) that there is a wealth of geo-tagged list data in Twitter; (ii) that these lists tend to focus on areas of potential expertise; and (iii) that distance impacts list labeling (and possibly revealing the localness of particular topics) – we turn in the next two sections to developing methods for identifying local experts.

Recall that we propose to measure a candidate v_i's local expertise by a combina-

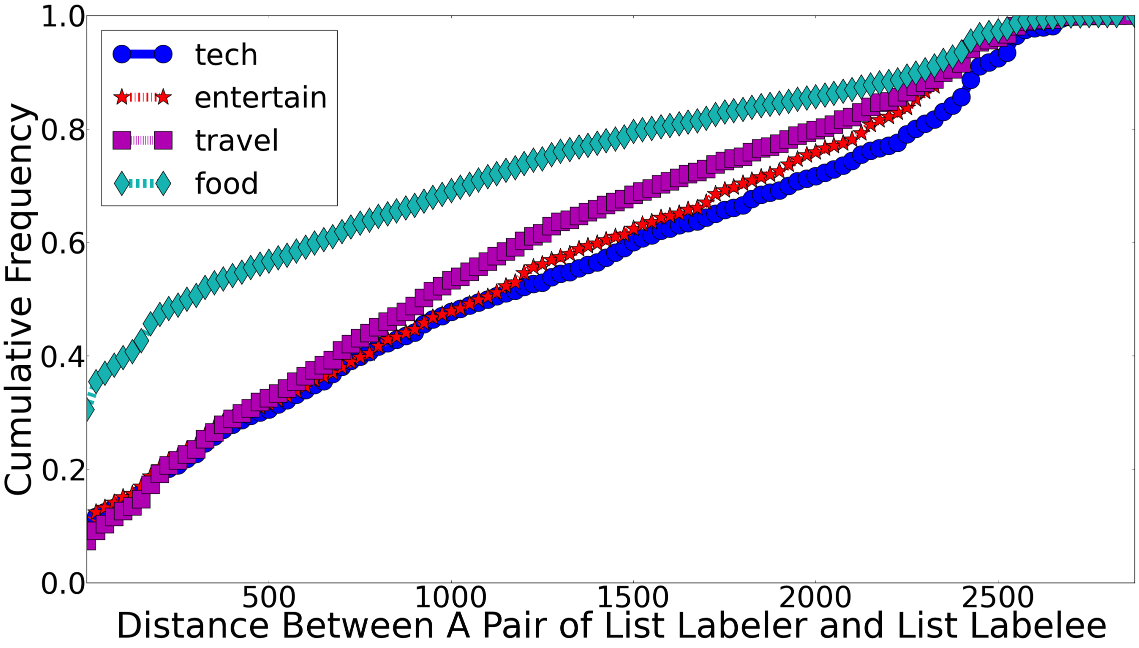
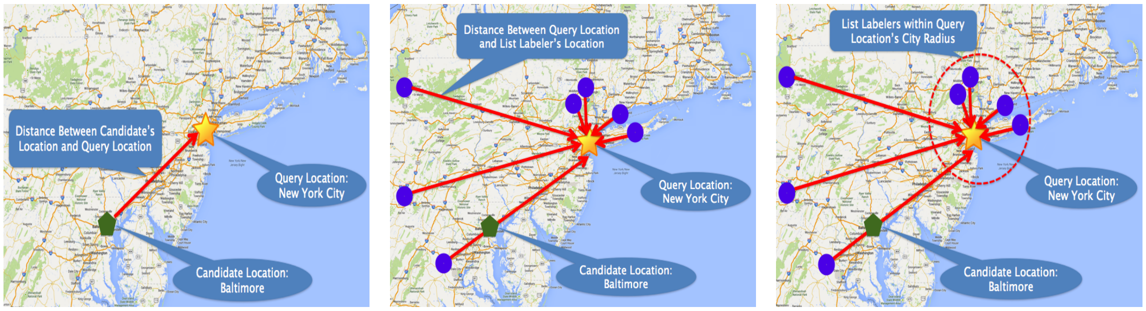


Figure 6.3: Cumulative Frequency of List Relationship Distances

tion of both the candidate's topical authority and local authority. While there are many ways to integrate these two scores, we propose a simple combination in this first study. We formally define candidate v_i's **LocalRank** (LR) $s(v_i, q)$ in query q as:

$$s(v_i, q) = s_l(l(v_i), l(q)) * s_t(\vec{v_i}, G, t(q))$$

where $s_l(l(v_i), l(q))$ denotes the *Local Authority* of v_i in query location $l(q)$, and $s_t(\vec{v_i}, G, t(q))$ denotes the *Topical Authority* of v_i in query topic $t(q)$ that is estimated using the candidate's expertise vector $\vec{v_i}$, and the social graph G that the candidate is involved in. In the following two sections we investigate how to estimate these values.


Distance Between Candidate's Location and Query Location
Query Location: New York City
Candidate Location: Baltimore
Distance Between Query Location and List Labeler's Location
Query Location: New York City
Candidate Location: Baltimore
List Labelers within Query Location's City Radius
Query Location: New York City
Candidate Location: Baltimore

(a) Candidate Proximity (b) Spread-based Proximity (c) Focus-based Proximity

Figure 6.4: Three Methods for Estimating Local Authority

6.4 Estimating Local Authority

In this section, we present three approaches for estimating a candidate expert's *local authority*. The first local authority method considers the distance between an expert candidate's location and the location of interest, capturing the intuition that closer candidates are more locally authoritative. The latter two approaches leverage the fine-grained geo-tagged linkages among the sampled Twitter users as revealed through list relationships, where both the list creator and the user being labeled have revealed a precise location.

6.4.1 Candidate Proximity

The first (and perhaps most intuitive) approach to estimate candidate v_i's local authority for query q is to use the distance between candidate v_i's location $l(v_i)$ and the query location $l(q)$. For example, if we are looking for experts on Austin Barbecue, then all candidates located in Austin will be considered more authoritative than candidates outside of Austin. We define this *Candidate Proximity* ($s_{l_{CP}}$) as:

$$s_{l_{CP}}(l(v_i), l(q)) = \left(\frac{d_{min}}{d(l(v_i), l(q)) + d_{min}} \right)^{\alpha}$$

where $d(l(v_i), l(q))$ denotes the distance between $l(v_i)$, and $l(q)$ (using the Haversine formula which accounts for the curvature of the earth), and we set $d_{min} = 100$ miles. In this case α indicates how fast the local authority of candidate v_i for query location $l(q)$ diminishes as the candidate moves farther away from the query location. This first local authority approach captures the intuition that closer candidates are more locally authoritative. Figure 6.4a shows a candidate expert in Baltimore (the green pentagon); if we are looking for an expert in New York (the gold star), such a Baltimore candidate's local expertise will be a function of the distance from Baltimore to New York. While simple, this approach cannot capture local expertise of candidates who do indeed live far from a location of interest. As we have mentioned before, Daniel Vaughn is an expert in Austin Barbecue although he lives 200 miles away in Dallas.[*]

To capture these cases where expertise is not dictated solely by distance from a candidate to an area of interest, we next propose two local authority methods that consider the distance of the candidate expert's "core audience" from the location of interest (that is, to reward candidates who have many labelers near the location of interest, even if the candidate lives far away).

6.4.2 Spread-based Proximity

The first of these geo-tagged list methods is the *Spread-based Proximity* that measures the "spread" of a candidate's core audience's locations compared to the query location:

$$s_{l_{SP}}(L(V_l(v_i)), l(q)) = \frac{\sum\limits_{v_{l_j} \in V_l(v_i)} s_{l_{CP}}(l(v_{l_j}), l(q))}{|V_l(v_i)|}$$

[*] In addition, the home location of an expert candidate may not even be accurate: recall that our home locator estimates a location based on a single user's geo-tagged tweets. In contrast, the following two local authority methods consider the aggregated perspectives of many list labelers, so there is a clearer signal of a candidate's location of expertise.

where v_{l_j} denotes one of the core audience $V_l(v_i)$ of candidate v_i. Basically, the "spread" it measures considers how far candidate v_i's core audience are from the query location $l(q)$ on average. If the core audience of a candidate is close to a query location on average, the candidate gets a high score of $s_{l_{SP}}$. For example, in Figure 6.4b, the green pentagon and the gold star represent the expert candidate's location and the query location, respectively. However, the spread-based proximity for the candidate in the query location emphasizes the distance of the links (plotted as red arrows) between the candidate's list labelers' locations (plotted as blue dots) and the query location.

6.4.3 Focus-based Proximity

In some cases, the spread-based proximity approach may underestimate a candidate's local authority. For example, for a couple of "foodies" v_a and v_b both in New York City, suppose v_a has a large audience in New York City recognizing his food expertise, and is well appreciated by a lot of people on the west coast, and even abroad; while v_b is much less well recognized by the local community in New York City, but has more people recognizing his expertise in mid-east United States, North Carolina, and Florida. Despite a much better local community recognition in New York, user v_a has a lower value of spread-based core audience query spatial proximity, due to the higher spatial spread of his labelers. To overcome this type of expertise underestimation, we propose the *Focus-based Proximity* as:

$$s_{l_{FP}}(L(V_l(v_i)), l(q)) = \frac{|\{v_{l_j} | d(l(v_{l_j}), l(q)) \leq r(l(q))\}|}{|V_l(v_i)|}$$

where $r(l(q))$ represents a radius around a location $l(q)$. This focus-based proximity measures how focused a candidate's audience is in the query location by measuring the percentage of the core audience that resides within the radius of the query location. For example, in Figure 6.4c, 4 out of 7 labelers (blue dots) for the candidate (green pentagons) are within the radius (plotted as the red dashed circle) of the query location (gold star), and the focus-based proximity in this case is $\frac{4}{7} \approx 0.57$.

These two local authority methods – the spread-based and focus-based approaches – are designed to capture the expert candidate's spatial influence measured via collective intelligence contributed by the people who labeled the candidate.

6.5 Estimating Topical Authority

In this section, we discuss how we estimate the topical authority score of candidate v_i being as a local expert in query q. Specifically, we propose to use both the crowd-sourced geo-tagged labels and the social connections between people to quantify a candidate's topical expertise score given a query.

6.5.1 Directly Labeled Expertise

We begin with a topical authority approach that leverages the directly labeled expertise of candidate v_i, as revealed through the sampled Twitter lists. Specifically, we adapt the user-centric model that Balog et al. proposed in [8] to estimate the *Topical Authority Score* $s_t(\vec{v_i}, G, t(q))$ of v_i with respect to the query topic $t(q)$ (ignoring for now the social graph G). Balog et al. applied the user-centric model to identify an expert's knowledge based on the documents (emails and web pages) that they are associated with. In our scenario, we apply the user-centric model to identify expert candidates' expertise based on the list labels that the crowd has applied to them.

The model is built on standard language modeling techniques: a user v_i can be represented by a multinomial probability distribution over the vocabulary of topic words (i.e., $p(t_w|\theta_{v_i})$, where θ_{v_i} denotes a user model). In this case, for each user v_i, we infer a user model θ_{v_i} such that the probability of a topic word t to occur in user

v_i's list labels can be estimated via $p(t_w|\theta_{v_i})$.

Given user v_i's user model θ_{v_i}, for a query q, user v_i's *Topical Authority Score* $s_t(\vec{v_i}, G, t(q))$ in query q is measured as the probability of query text $t(q)$ to be generated from the users' user model:

$$s_t(\vec{v_i}, G, t(q)) = p(t(q)|\theta_{v_i}) = \prod_{t_w \in t(q)} p(t_w|\theta_{v_i})$$

where t_w denotes a topic word in query text $t(q)$.

Since we are expecting that most of the users will be labeled by a small number of unique labels, most of the topic words will have zero probabilities for a particular user v_i. Thus we smooth $p(t_w|\theta_{v_i})$ using the probability of the topic word to occur in the whole corpus of labels $p(t_w|\theta_{C_v})$ when estimating $p(t_w|\theta_{v_i})$:

$$p'(t_w|\theta_{v_i}) = (1 - \lambda) * p(t_w|\theta_{v_i}) + \lambda * p(t_w|\theta_{C_v})$$

Here λ represents the extent of smoothing. A large value of λ indicates that the probability $p(t_w|\theta_{v_i})$ is more weighted towards the probability of the topic word t_w to occur in the corpus $p(t_w|\theta_{C_v})$. In the experiments, we fix the value of λ to 0.1.

6.5.2 *Expertise Propagation*

In addition to the directly labeled expertise derived from our collection of geo-tagged Twitter lists, we are interested to explore whether the social and list-based connections of Twitter users also provide strong signals of expertise. Specifically, we consider three graphs that include three types of connections: (i) User Friendship; (ii) List-labeling Relationship; and (iii) List-peer Relationship (see the data collection described in Table 6.1). Recall that each user v_i is characterized as a vector $\vec{v_i}$ of his topical expertise generated from the directly labeled expertise method. Can we

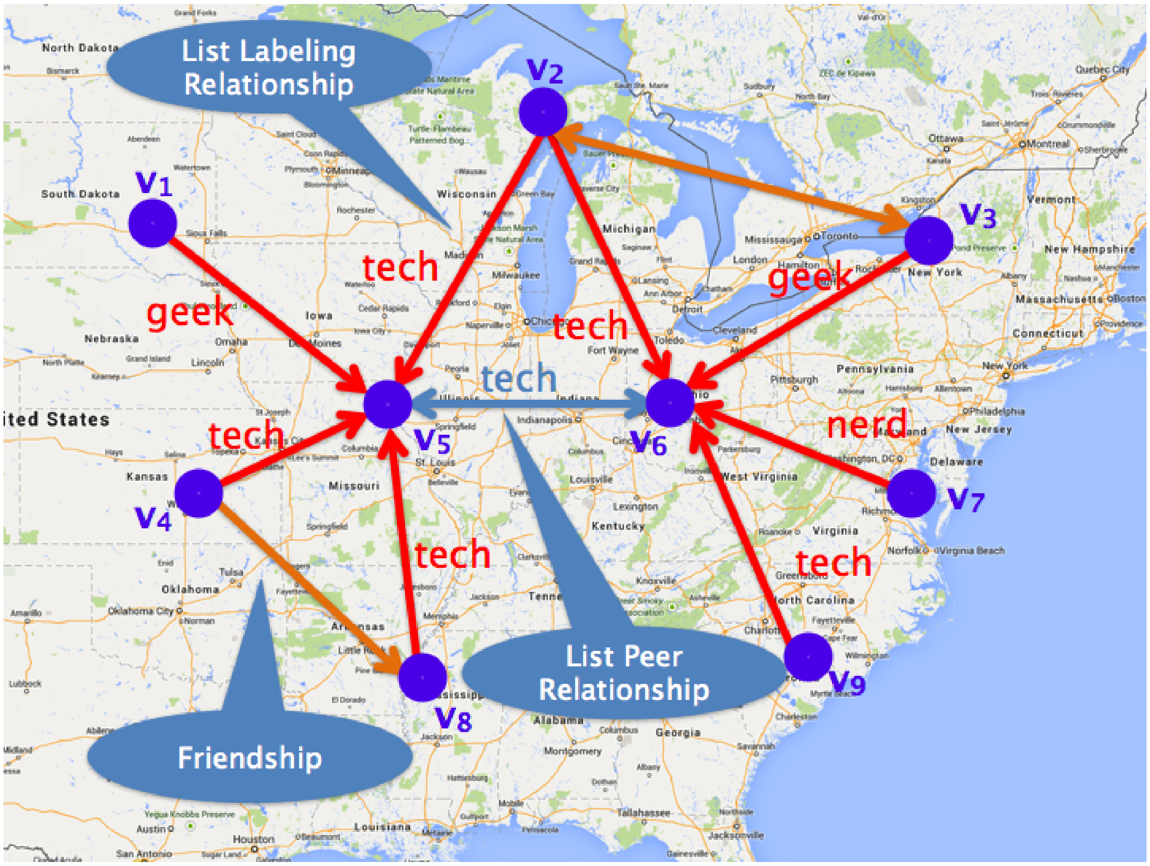


Figure 6.5: Examples of Social and List-based Connections

enrich the expertise signals from the Twitter lists by propagating expertise along these three graphs? The intuition is that people with particular expertise have a higher likelihood to be connected to other people with the same expertise.

6.5.2.1 User Friendship

The first expertise propagation approach is based on user friendship, as represented by a direct link $e(v_i, v_j)$ from user v_i to user v_j. In Figure 6.5, we show nine expert candidates (plotted as blue dots that are labeled from v_1 to v_9). Here, a *friendship link* (plotted as an orange arrow) connects a candidate to another candidate that he follows, and an example would be the orange arrow on the bottom left from v_4 to v_8. The motivation for propagation along friendship links is that a candidate has a higher likelihood to be an expert in query topic $t(q)$ if he has friend(s) that are also expert(s) in query topic $t(q)$.

Given users' friendship linkages, we can generate the friendship graph $G_f(V, E)$ for a set of users V, and a set of friendship links E that connect users in V. For every edge $e_f(v_i, v_j)$, the weight $w_f(v_i, v_j)$ is simply $\frac{1}{|E_{out}(v_i)|}$, where $E_{out}(v_i)$ represents the set of out links from user v_i.

In addition, from the perspective of the "First Law of Geography" [113] that "everything is related to everything else, but near things are more related than distant things", we hypothesize that a user knows a friend nearby better than a friend farther away. Thus, we generate an alternative ${G_f}'(V, E)$ to reflect the effect of distance between a pair of connected users v_i, and v_j on how well user v_i knows v_j (i.e., how much credit v_j gets from v_i), by introducing the local authority score to the calculation of the weight ${w_f}'(v_i, v_j)$ for edge $e(v_i, v_j)$:

$${w_f}'(v_i, v_j) = \frac{s_l(l(v_i), l(v_j))}{|E(v_i)|}$$

6.5.2.2 List-labeling Relationship

The second expertise propagation approach considers the *list-labeling relationship* derived from the sampled Twitter lists. The motivation for the propagation here is: if an expert v_i in a topic $t(q)$ labels another user v_j as an expert in the same topic, user v_j also has a high likelihood to be an expert in the topic.

For example, user v_i lists user v_j as a tech expert in one of his lists on Twitter, generating a direct link $e_l(v_i, v_j)$ from v_i to v_j indicating a relationship connected by expertise recognition. In this way, a graph G_l capturing the expertise recognition can be constructed. Returning to Figure 6.5, we show this list-labeling relationship (plotted as a red arrow) that links a list labeler to the candidate that he listed, and an example would be the red arrow on the top left from v_1 to v_2 with a list label "geek".

As in the friendship case, we can similarly construct two graphs – one with the weight $w_l(v_i, v_j)$ and the other one with the distance-based weight $w_l{}'(v_i, v_j)$ for the link $e_l(v_i, v_j)$ according to the number of out links from v_i in G_l, and $G_l{}'$ respectively.

6.5.2.3 List-peer Relationship

Finally, we can propagate expertise along peers that appear on the same list. Returning to Figure 6.5, this list-peer relationship (plotted as a blue arrow) indicates a connection between two candidates that appear on the same list, and examples in the figure are the blue arrows in the middle between v_5 and v_6 with a list label "tech". The list peer relationship carries an important signal to infer a list member's expertise for being peers on the same list with top experts. For example, a user who co-occurs with several top tech experts on lists also has a good chance to be a tech expert.

Here, we have the link $e_{lp}(v_i, v_j)$ that directly connects user v_i to user v_j in a list on Twitter. We can measure the weight $w_{lp}(v_i, v_j)$ for the link $e_{lp}(v_i, v_j)$ according to the number of out links from v_i in G_{lp}. Using all the list peer relationship, we generate a social graph G_{lp} that captures the signals of expertise propagated from list peers. We can also generate the corresponding distance-weighted list peer graph $G_{lp}{}'$.

6.5.2.4 Topical Authority Score from Expertise Propagation

Given these three perspectives, we propagate expertise along these graphs through a random walk based on topic-sensitive PageRank (TSPR) [49]. Again, our intuition is that people with particular expertise have higher likelihood to be connected to other people with the same expertise. The random walk approach leverages this intuition by propagating expertise along links in the graph, and by resetting back to the candidates with high directly labeled expertise. Thus, for each particular social graph G described above (that is: G_f / $G_f{}'$, G_l / $G_l{}'$, or G_{lp} / $G_{lp}{}'$), we apply TSPR on the specific social graph to identify the most influential users for a particular query topic $t(q)$. The stabilized TSPR score for each user v_i is considered as user v_i's topical authority score $s_t(\vec{v_i}, G, t(q))$ in query topic $t(q)$. In our experiments, we explore using both the general social graph, and the distance weighted social graph to identify top local topical experts for a given query.

6.6 Evaluation

In this section, we evaluate the proposed local expert finding framework. We seek answers to the following questions:

- What impact does the choice of local authority have on the quality of local expert finding in LocalRank? How much do crowdsourced geo-tagged list labels impact local authority (and ultimately the quality local expert finding)?

- Do the three types of expertise propagation over social and list-based connections of Twitter users provide strong signals of topical expertise? And if so, to what degree over directly labeled expertise?

- How well does LocalRank perform compared to alternative local expert finding approaches? Is integrating topical and local authority necessary?

- Finally, how do the approaches perform in finding top local experts for finer topics? Do we see consistent performance in comparison with more general topics?

6.6.1 Experimental Setup

In this subsection, we first describe the location + topic queries and then introduce the specific expert finding approaches we tested. We discuss how we gathered ground truth to evaluate these approaches, and how we measured approach effectiveness.

6.6.1.1 Queries

In total, we evaluate local expert finding using 56 queries (16 general topic queries and 40 finer topic queries). We consider four general query topics coupled with four locations, totaling 16 topic-location queries. Specifically, we look for local experts in the areas of "technology", "entertainment", "food", and "travel" in New York City, Houston, San Francisco, and Chicago. We also consider 10 refined topics under the general umbrella of "food" and "startup", again in the same locations, totaling 40 topic-location queries. These refined topics are "barbecue", "seafood", "pizza", "winery", and "brewery" under the "food" scenario, and "venture capital", "incubator", "founder", "entrepreneur", and "angel investor" under the "startup" scenario. By considering both general-topic and finer-topic local expertise queries, our goal is to investigate differences in local expertise finding at varying granularities of expertise.

6.6.1.2 Approaches for Finding Local Experts

In addition to the proposed local expert finding approaches presented in this section, we consider five alternative baselines. The first considers only a candidate's topical authority (ignoring local authority):

- *Directly Labeled Expertise* (*DLE*): Rank candidates by topical authority in the query topic.

The next three consider only a candidate's local authority (ignoring topical authority):

- *Nearest* (*NE*): Rank candidates by distance to the query location.
- *Most Popular in Town by Followers Count* (*MP (follower)*): Rank candidates from the query location by follower count.
- *Most Popular in Town by Listed Count* (*MP (list)*): Rank candidates from the query location by the number of lists the candidate appears on.

The final baseline combines simple versions of topical and local authority:

- *Most Popular in Town by Listed Count on Topic* (*MP (on-topic)*): Rank candidates from the query location by the number of on-topic lists the candidate appears on.

We compare these five baselines with the proposed LocalRank approach presented in this section. For LocalRank, we investigate the three approaches for estimating local authority – by Candidate Proximity (CP), Spread-based Proximity (SP), and Focus-based Proximity (FP) – and the Directly Labeled Expertise (DLE) and Expertise Propagation (EP) approaches for estimating topical authority. When applying both the *Candidate Proximity*, and *Spread-based Proximity*, we preset the d_{min} to be 100 (miles), and *alpha* to be 2.0. We calculate the local expertise score using the normalized topical authority score and the normalized local authority score.

6.6.1.3 Gathering Ground Truth

Since there is no explicit data that directly specifies a user's local expertise given a query (location + topic), we gather ground truth by employing human raters on Amazon Mechanical Turk. For each of the experimental settings (an approach

+ a query topic + a query location), we retrieve the corresponding top-10 local expert candidates with the highest local expertise scores, and have human raters on Mechanical Turk label to what extent an expert candidate has local expertise in the query topic and the query location. For each expert candidate, 5 turkers (human raters) label the candidate's local expertise using a four-scale local expertise rating:

- Extensive Local Expertise [+2]: The candidate has extensive expertise in the query topic, and is locally well recognized in the query location for his expertise.
- Some Local Expertise [+1]: The candidate has some expertise in the query topic, and also has some influence in the query location
- No Evidence [0]: The candidate has no clear evidence to be considered as having expertise in the query topic, or influence in the query location.
- No Local Expertise [-1]: The candidate has neither any expertise in the query topic, nor influence in the query location.

For each assessment, we provide the turker with the candidate's user profile, a word cloud generated using the labels that people used to describe the candidate, a heatmap showing the locations of the candidate's labelers, the candidate's most retweeted 5 tweets and 5 most recent tweets. To ensure the quality of these assessments, we follow the conventions suggested by Marshall and Shipman [82]. Each individual HIT (Human Intelligence Task) includes 10 query / expert candidate pairs randomly selected from all the pairs of query and expert candidate. 2 out of the 10 pairs for each HIT are manually labeled by domain experts in order to evaluate the quality of the feedback from turkers. If a turker picks a significantly different answer comparing to ours for either one of the two particular pairs, the feedback for the HIT will be discarded. For a particular pair of query and expert candidate, we use the best judgment (i.e., the most voted rating) out of the 5 turkers as the final rating for the pair.

We investigate the inter-judge agreement using both *kappa statistic* and *Accuracy*. Since we have more than two annotators (five in our scenario) for each query-candidate pair, we adopt Fleiss' kappa [38], which ranges from 0 (when the agreement is not better than chance) to 1 (when the two annotators agree with each other perfectly). Following Brants [12] and Nowak et al. [89], we define *Accuracy* as:

$$Accuracy(Q_{pairs}) = \frac{\sum_{q_{pair} \in Q_{pairs}} \frac{\text{\# of votes for the majority rating}}{\text{\# of votes for } q_{pair}}}{|Q_{pairs}|}$$

where Q_{pairs} represents the set of query and candidate pairs, in which each pair q_{pair} includes both a query q, and an expert candidate c. An ideal *Accuracy* would be 1.0 that all the turkers pick the same local expertise rating for every particular pair of query and candidate. For example, an *Accuracy* of 0.6 indicates that for a query-candidate pair, 60% of the human raters voted for the majority choice.

6.6.1.4 Metrics

To evaluate each local expert finding approach, we measure the average *Rating@k*, *Precision@k*, and *NDCG@k* across all queries in our testbed. For the following experiments, we consider all the 0 and -1 ratings as 0s.

Rating@k measures the average local expertise ratings by the human-raters for the top k ranked local experts across all the queries:

$$Rating@k = \frac{\sum_{q \in Q_{pairs}} (\sum_{i=1}^{k} rating(c_i, q)/k)}{|Q_{pairs}|}$$

where Q_{pairs} represents the set of all query pairs, and $rating(c_i, q)$ denotes the most voted local expertise rating for candidate c_i in query q. *Rating@k* ranges between 0 to 2, and an ideal approach will have a *Rating@k* value 2, which all identifies local experts with extensive local expertise in the query topics and locations. Conversely, the worst performing approach will have a *Rating*@k value 0, indicating that the approach only identifies local experts as those with no local expertise or no evidence.

Precision@k measures the average percentage of candidates that are relevant to the query topic and query location in the top k candidates across all the queries. It is defined as:

$$Precision@k = \frac{\sum\limits_{q \in Q_{pairs}} \frac{|\{c_i | rating(c_i, q) > 0\}|}{k}}{|Q_{pairs}|}$$

In this section, we consider expert candidates with both "extensive local expertise", and "some local expertise" as relevant, while we consider both "no local expertise" and "no evidence" as irrelevant. A perfect local expertise estimator has a $Precision@k$ value of 1.0.

NDCG@k (Normalized Discounted Cumulative Gain@k) measures how well the predicted local expert rank order is compared to the ideal rank order (i.e., candidates are ranked according to their actual local expertise) for the top k results across all the query pairs. $NDCG@k$ ranges between 0 and 1, and a higher value indicates an approach that generates better rank orders.

6.6.2 Agreement of Local Expertise

Before evaluating the proposed local expert finding framework, we are interested to study how consistent and reliable the results from Mechanical Turk are. Overall, we have 11,285 individual judgments made by the human raters. Is local expertise discernible? And is local expertise assessment consistent across topics and locations?

To start with, we report the kappa (κ) and *Accuracy* values in Table 6.3. When considering 3 rating categories for each pair (2: "extensive local expertise", 1: "some local expertise", and 0: either "no local expertise" or "no evidence"), the overall

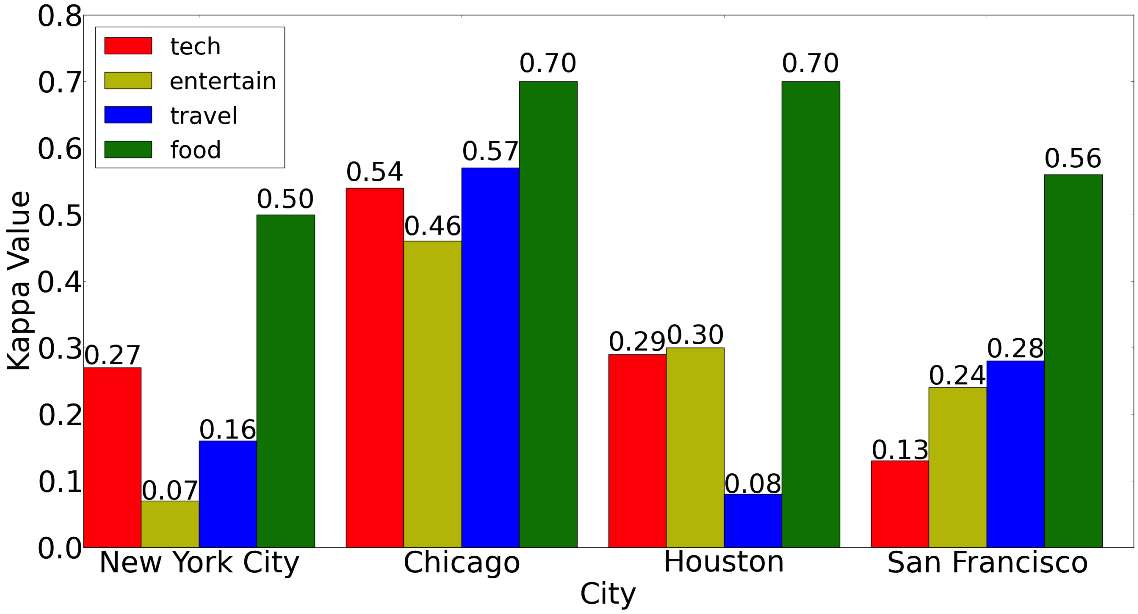


Figure 6.6: Kappa Value by Query for Binary Rating Categories

Table 6.3: User Agreement for Overall Judgments

	3 Rating Categories		2 Rating Categories	
Overall	Accuracy	κ	Accuracy	κ
	0.716	0.280	0.822	0.397
General Topics	Accuracy	κ	Accuracy	κ
	0.715	0.279	0.818	0.393
Finer Topics	Accuracy	κ	Accuracy	κ
	0.717	0.281	0.825	0.401

Accuracy for agreement is 0.716, indicating that for a pair of query and candidate, on average 71.6% of the human raters voted for the majority vote. This demonstrates good user agreement and is significantly higher than accuracy by chance (33.3% for three categories). When considering only 2 rating categories (2 and 1 as relevant, and 0 as irrelevant), the overall *Accuracy* increases to 82.2%, which is also much higher than the accuracy by chance (50% for two categories). For kappa, we see that the overall value is 0.280 in the 3 rating category case. For the binary rating case, the overall kappa value is 0.397. Both kappa statistics are typically considered "fair" inter-judge agreements. Together, these kappa and *Accuracy* values suggest that these human raters have a fairly good agreement. And we observe both much higher *Accuracy* and kappa value for binary rating categories, which indicates that raters find it easier to decide whether a candidate has local expertise or not, rather than determining the extent of a candidate's local expertise.

Decomposing expertise into our four locations of interest (NYC, Chicago, Houston, and San Francisco), we see in Figure 6.6 that inter-judge agreement varies greatly by both location and by topic. Some locations are much easier to assess (Chicago) than other locations (NYC). Looking into the top local experts identified for New York City, we observe that some of them are national celebrities from NYC, which makes it trickier for human raters to decide whether they are really local experts or not. Confirming our observation of "food" being a more local topic as revealed through Twitter lists in Figure 6.3, we see that local "food" expertise is easier to agree upon, whereas other topics are more difficult. For the candidate and query pairs for finer topics, we observe slightly higher values of *Accuracy* and kappa. In terms of kappa values for particular query topics, "angel investor" and "brewery" have the highest kappa value of 0.461, and 0.532 for "startup" and "food" scenario. These results show that some topics are inherently more local, and thus could be easier for human raters to judge the expertise of candidates in those topics.

6.6.3 Comparing Local Expert Finding Approaches

In this subsection, we seek answers for the questions brought up in the beginning of this section, with four set of experiments: (i) evaluating the performance of local

Table 6.4: LocalRank: Evaluating the Three Local Authority Approaches

Local Authority	*Rating*@10	*Precision*@10	*NDCG*@10
CP	0.952	0.553	0.685
SP	1.330	0.830	**0.903**
FP	**1.334**	**0.842**	0.896

authority metrics; (ii) studying the impacts of expertise propagation; (iii) comparing the performance of baseline approaches and the LocalRank approaches; and (iv) evaluating the performance of expert finding via finer topics.

6.6.3.1 LocalRank: Evaluating Local Authority

To begin with, we seek to understand the impact of the local authority approach on the quality of local expert finding in LocalRank. Specifically, we fix the LocalRank topical authority as the Directly Labeled Expertise, while we vary the local authority across the three approaches presented in Section 6.4: Candidate Proximity (CP), Spread-based Proximity (SP), and Focus-based Proximity (FP). Our goal is to understand to what degree the local authority affects local expert finding, and to assess if (and how much) the crowdsourced geo-tagged list labels impact local authority.

We present in Table 6.4 the *Rating*@10, *Precision*@10, and *NDCG*@10 for each of the three local authority approaches. We observe that both of the approaches (SP and FP) that utilize the locations from the candidates' core audience significantly improve the performance of local expert finding in comparison with the candidate proximity approach (CP) that only takes the candidate's physical location into consideration. Using candidate proximity (CP), the LocalRank approach only identifies true local expert 55% of the time on average among the top 10 candidates. Similarly, we see comparatively low values of *Rating*@10 as 0.952, and $NDCG$@10 as 0.685. In contrast, the Spread-based Proximity (SP) and Focus-based Proximity (FP) approaches reach $Precision@k$ of almost 85%, *Rating*@10 over 1.33, and $NDCG$@10 of 0.90. This indicates the core audience for an expert candidate is crucial to estimating a candidate's local authority. And in absolute terms, the rating scores for both approaches range between "some local expertise" (1) and "extensive local expertise" (2), indicating that these approaches can identify candidates who are actually local experts. Interestingly, we see for this evaluation framework that the two approaches perform nearly equally well, although they capture two different perspectives on local authority (recall that SP considers the average distance of labelers, whereas FP considers the fraction of labelers within a radius).

6.6.3.2 LocalRank: Impact of Expertise Propagation

Given these results for local authority, we next consider the impact of expertise propagation on the topical authority (and ultimately on the quality of local expert finding). As described in Section 6.3.2, we explore whether the three types of social and list-based connections of Twitter users do indeed provide strong signals of expertise. We consider the (i) friendship graph, (ii) list-labeling relationship graph, and (iii) list-peer relationship graph. For each graph (both with and without distance-weighted edges), we apply the topic-sensitive PageRank algorithm to propagate expertise. For each particular graph as well as a particular type of edge weight, we iterate the damping factor from 0.10 to 0.30 to 0.50 to study how the damping factor affects the task of finding top local experts. A smaller damping factor indicates less score propagation and more random walking among more topic-relevant nodes in the graph. We find that the conventional damping factor value (0.85 or 0.90) finds only national celebrities like @JimmyFallon (Jimmy Fallon, host of talk

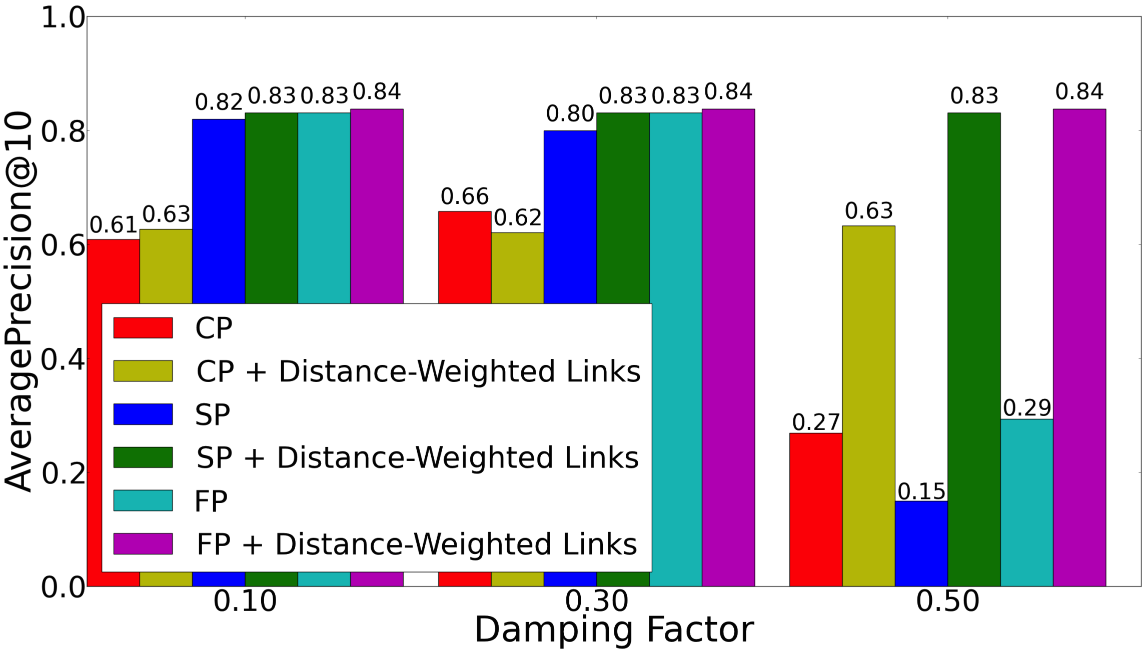


Figure 6.7: *Precision*@10 with Friendship as Input Graph

show Late Night with Jimmy Fallon), @TheEllenShow (Ellen Degeneres, host of the Ellen Degeneres Show), and @Jack (Jack Dorsey, Twitter and Square co-founder) no matter what the query topic is. With a smaller value of damping factor, we hope to identify more topical relevant local experts.

We present in Figure 6.7 the local expertise results for expertise propagation using the Friendship graph as input, coupled with corresponding parameter settings. We vary the choice of local authority (CP, SP, and FP), the use of distance-weighted links or not, as well as the choice of damping factor. This figure focuses on *Precision*@10, while the subsequent Table 6.6 in p. 165 includes *Rating*@10, *Precision*@10, and *NDCG*@10 for all graph types. First, in terms of the damping factors, we see that across all settings (0.10, 0.30, and 0.50), that the best performing result is comparable. However, we do observe a significant performance drop for damping factor 0.50 using regular edge weight that does not consider distance between the nodes as a factor. Upon investigation into the top local expert candidate under this setting, we observe that many of the top local candidates are national celebrities (e.g., @JimmyFallon, @TheEllenShow, and @Jack), compared to the candidates retrieved using a damping factor of 0.10 or 0.30. We attribute this result to the higher weight on score propagation through general friendship edges. On the other hand, for a damping factor 0.10 or 0.30, most of the scores are propagated through topic-relevant nodes via random walking.

Second, we observe a slight improvement for distance-based edge weight when using a damping factor of 0.10 or 0.30 rather than using the regular edge weight. And we observe a dramatic improvement of performance for distance-weighted edge weight using a damping factor of 0.50 than the alternative version. This indicates that giving local friends more credit (in terms of expertise propagation flowing more strongly to nearby friends than far away ones) does help improve the likelihood to find better top local experts.

Third, in terms of the choice of location authority metric, we observe a similar result to what we observed in the previous section – that the approaches (SP and FP) that utilize the locations from the candidates' core audience significantly improve the performance of local expert finding.

Finally, compared to the simpler approach of not propagating expertise at all, but just using the directly labeled expertise, we see that the results are quite similar (with $Precision$@10 near 0.84). Given this result, we compared the lists of top-10 local experts returned by LocalRank using directly labeled expertise versus LocalRank using each one of the expertise propagation approaches. While the overall precision is similar, the experts that each approach finds are different: we find an average Jaccard coefficient between local expert lists of around 60 to 80%. In other words, on average, 20 to 40% of the top-10 local experts for the same query are different, when we compare the directly labeled expertise approach versus a particular expertise propagation approach. This indicates that the expertise propagation approaches are bringing in new signals of local expertise from the social and link-based connections of users; in our continuing work we are investigating methods to integrate these two types of topical authority by finding more diverse experts from each of these alternative approaches.

Table 6.5: Comparing LocalRank to Five Alternatives

Approach	*Rating*@10	*Precision*@10	*NDCG*@10
DLE	0.225	0.088	0.199
NE	0.141	0.114	0.487
MP (followers)	0.058	0.031	0.234
MP (list)	0.070	0.038	0.301
MP (on-topic)	1.059	0.628	0.750
LR: SP + DLE	1.334	**0.842**	**0.896**
LR: SP + EP + Friendship	**1.354**	0.838	0.884

6.6.3.3 Comparing LocalRank versus Alternatives

So far we have investigated the impact of local authority and the impact of topical authority on the quality of local experts found by the LocalRank framework. In this section, we compare LocalRank to the five alternative local expert finding approaches described in the experimental setup over the set of 10 general topics.

We first report the results for the five baselines in Table 6.5. We see that relying solely on topical authority – Directly Labeled Expertise (DLE) – with no notion of localness, results in a very low *Rating*@10, *Precision*@10, and *NDCG*@10. Similarly, relying solely on local authority – Nearest (NE), Most Popular in Town by Followers Count (MP followers), and Most Popular in Town by Listed Count (MP list) – with no notion of topical authority also leads to very poor results. Since local experts are defined both by their localness and their on-topic expertise, these results confirm our intuition driving the LocalRank approach to combine both factors. The baseline that does incorporate both factors – Most Popular in Town by Listed Count on Topic (MP (on-topic)) – captures this notion of local expertise by rewarding candidates who have been listed on many Twitter lists on the topic of interest within a particular location. We see in the table that this approach significantly outperforms the single factor alternatives (*Rating*@10 of 1.059, *Precision*@10 of 0.628, and *NDCG*@10 of 0.750).

Table 6.6: The Impact of Expertise Propagation on LocalRank versus the Best Performing Alternative (% of Imp: % of Improvement)

Approach	*Rating*@10	% of Imp	*Precision*@10	% of Imp	*NDCG*@10	% of Imp
MP (on-topic)	1.059	–	0.628	–	0.750	–
LR: DLE + Local Authority	1.334	26.0%	0.841	33.9%	**0.897**	**19.6%**
LR: EP + Friendship Graph	1.354	27.6%	0.838	33.4%	0.884	17.9%
LR: EP + List-labeling Graph	**1.354**	**27.6%**	**0.847**	**34.9%**	0.886	18.1%
LR: EP + List-peer Graph	1.345	27.0%	0.844	34.4%	0.887	18.3%

We compare all five of these baselines to two versions of LocalRank. Both consider local authority based on Spread-based Proximity (SP); one uses directly labeled expertise (SP + DLE), while the other uses expertise propagation (SP + EP + Friendship) over the friendship graph. We see similar qualitative results when evaluating Focus Proximity (FP) and alternative expertise propagation approaches.

Both approaches significantly outperform the four single factor baselines, as well as significantly outperforming the best alternative incorporating both local and topical authority, MP (on-topic). We see for LocalRank (SP + DLE) a *Rating*@10 of 1.334, *Precision*@10 of 0.842, and *NDCG*@10 of 0.896. For LocalRank (SP + EP + Friendship), we have *Rating*@10 of 1.354, *Precision*@10 of 0.838, and *NDCG*@10 of 0.884. These results confirm the effectiveness of the LocalRank approach and the importance of carefully leveraging the large-scale geo-tagged list relationships on Twitter.

Continuing this investigation, we report the results of the different LocalRank approaches versus the best performing baseline in in Table 6.6. We see that the Expertise Propagation approaches generally perform slightly better than the Directly Labeled Expertise approach in terms of *Rating*@10 and *Precision*@10. This suggests that adding in social connections bring in more signals to identify top local experts. In particular, LocalRank with expertise propagation coupled with the social graph of list-labeling relationships generates the best performance, with *Rating*@10 of 1.354 (an improvement of 27.6% over MP (on-topic)), *Precision*@10 of 0.847 (an improvement of 34.9%), and *NDCG*@10 of 0.886 (an improvement of 18.1%). However, in terms of *NDCG*@10, we see that the simpler DLE approach performs slightly better. But in all cases, the LocalRank approach outperforms the alternative.

6.6.3.4 LocalRank: Local Experts Over Finer Topics

Finally, we drill down from general topics to more fine-grained topics, to investigate the ability of local expertise finding approaches to handle these more specific cases. Here we evaluate the proposed LocalRank approaches via the refined topics under the “food”, and “startup” scenarios. We report the performance using the best parameter settings for each of the proposed approaches. In this experiment, we

Table 6.7: Comparing LocalRank versus the Best Performing Alternative over Finer Topics

Approach	*Rating*@10	*Precision*@10	*NDCG*@10
MP (on-topic)	0.782	0.526	0.707
LR: SP + DLE	**0.924**	**0.583**	**0.851**
LR: SP + EP + Friendship	0.871	0.538	0.846
LR: SP + EP + List-labeling	0.868	0.535	0.837
LR: SP + EP + List-peer	0.865	0.533	0.844

set local authority as using Spread Proximity and expertise propagation relies on a damping factor of 0.30.

Table 6.8: How Well does LocalRank Perform on Finer Topics?

Query Topic	*Rating*@10	*Precision*@10	*NDCG*@10
barbecue	0.631	0.404	0.787
seafood	0.825	0.525	0.868
pizza	0.775	0.425	0.712
brewery	**1.178**	**0.738**	**0.928**
winery	0.763	0.475	0.744
entrepreneur	**1.248**	**0.800**	**0.921**
venture capital	1.180	0.663	0.956
angel investor	0.923	0.638	0.846
incubator	0.660	0.413	0.732
founder	0.995	0.688	0.786

Table 6.7 presents the local expert finding results for the four types of LocalRank versus the best performing alternative (MP (on-topic)). We observe that once again the LocalRank approaches outperform the best-performing alternative in all cases.

However, we notice that the performance for these finer topics is worse than what we observed for the more general topics. For example, LocalRank with Directly Labeled Expertise performs the best with *Rating*@10 of 0.924, *Precision*@10 of 0.583, and *NDCG*@10 of 0.851 over these finer topics. But the same approach over the more general topics results in an average *Rating*@10 of nearly 0.4 points higher. Similarly, we see improved performance over the other metrics in the general topic case. We believe these results reflect two challenges: (i) First, it is fundamentally more challenging to identify local experts for more refined topics. For example, it may be easier to assess whether someone is a "food" expert, rather than that they are an expert in a specific topic like "barbecue". (ii) Second, there is inherent data sparsity at the level of these finer topics. The number of candidates for a finer topic in a query location is much smaller compared to the number of candidates for a general topic in the same query location. For example, we observe that the approaches consider the probable No. 1 barbecue expert in Texas – Daniel Vaughn – as a local expert for barbecue for query locations of Chicago and San Francisco, in addition to his natural expertise in Houston. For these two distant locations, Vaughn is often a top choice since there are few barbecue candidates recognized in the location.

In our continuing work, we are investigating the contours of expertise across the country, so that topics with a strong regional factor (like Barbecue, with its traditional centers in Texas, North Carolina, and the Midwest) can be balanced with topics of expertise that are found nearly everywhere (e.g., the more general "foodies"). Along these lines, we show in Table 6.8 the results of LocalRank (SP + DLE) for each of the fine-grained topics. As we observed in our original investigation of Twitter lists, where we observed topics like "food" being more local than topics like "technology", here we see great variation in local expertise finding across these different subtopics.

6.7 Summary

The exponential growth in social media over the past decade has recently been joined by the rise of location as a central organizing theme of how users engage with online information services and with each other. Enabled by the widespread adoption of GPS-enabled smartphones, users are now forming a comprehensive geo-social overlay of the physical environment of the planet. In this section, we have argued for leveraging these geo-spatial clues embedded in Twitter lists to power new local expert finding approaches. We have proposed and evaluated the LocalRank framework for finding local experts, by integrating both a candidate's local authority and topical authority. We have seen that assessing local authority based on the spread and focus-based proximity of a candidate's "core audience" – that is, the users who have labeled him – can lead to good estimates of local authority and ultimately to high-quality local expert finding. Through an investigation of 56 queries coupled with over 11,000 individual judgments from Amazon Mechanical Turk, we have seen high average precision, rating, and NDCG in comparison with alternatives. In our continuing work, we are interested to (i) further investigate the borders of "localness" by investigating when an expert is considered a local expert versus a regional expert; (ii) enhance our current LocalRank approach with temporal signals to capture expertise evolution; and (iii) incorporate the detected local experts into a prototype system that can direct information needs to local experts who are considered authoritative and responsive on the local topic of interest.

7. CONCLUSIONS AND FUTURE DIRECTIONS

We believe that the increasing ubiquity of location-based social media has the potential to fundamentally disrupt basic scientific inquiry into questions that heretofore were difficult to study and to provide the basis for new "intelligent" geo-social information systems. Accomplishing this will require new methods, new algorithms, and new frameworks for mining and analyzing vast fine-grained (public) spatio-temporal footprints, as well as new systems and techniques to leverage these footprints. We have outlined some of the challenges facing this opportunity and highlighted five of our related efforts toward informing this emerging research area. Moving forward, we believe that geo-social intelligence research is poised to make major breakthroughs in the years to come due to the growing interests of social scientists in computational/data-intensive approaches and the 4th paradigm [67, 54] and computer scientists in spatial computing [25]. We also believe that transformative research in geo-social system can be accelerated along multiple fronts if we continue to embrace and fine-tune the emerging open science paradigm [111] to promote interdisciplinary collaboration and improve the infrastructure for geo-social intelligence research.

7.1 Conclusions

Specifically, in this dissertation research, we focused on investigating the real-time geo-social footprints that connect people's online presence to their activities in the physical world. These real-time geo-social footprints correspond to the check-ins on Foursquare or Google Local, geo-tagged postings, conversations, and social connections on Facebook or Twitter, and so on. The access to the geo-social footprints of hundreds of millions of people brings in unprecedented opportunities of deeper and more insightful geospatial understanding of the emergent collective knowledge embedded in these geo-social footprints, and furthermore building new geo-social information systems utilizing these geo-social footprints. However, there is still a significant gap toward understanding, and leveraging these geo-social footprints. Thus, this dissertation made contributions towards two general directions:

First, we investigated the capacity of using these geo-social footprints to build new geo-social information systems. Specifically, we tackled the challenges of location sparsity, lack of understanding of these geo-social footprints, and lack of understanding of the viability of the public-shared geo-social footprints to complement and even replace traditionally more expensive proprietary data. (i): in order to tackle the lack of user adoption of geo-based features per user or per post signals, we proposed and evaluated a probabilistic framework for estimating a microblog user's location based purely on the content of the user's posts. With the help of a classification component for automatically identifying words in tweets with a strong local geo-scope, and a lattice-based neighborhood smoothing model for refining a user's location estimate, we have seen how the location estimator can place 51% of Twitter users within 100 miles of their actual location. (ii): to have a better understanding of the newly emerged location sharing services, we investigated a set of 22 million check-ins across 220,000 users and reported a quantitative assessment of human mobility patterns by analyzing the spatial, temporal, social, and textual aspects associated with these footprints. Concretely, we observe: (a) users follow simple reproducible mobility patterns; (b) social status, in addition to geographic and economic factors, is coupled with mobility; and (c) content and sentiment-based analysis of posts can reveal heretofore unobserved context between people and locations. (iii): to verify whether publicly-shared data is viable to capture real-world flows of people instead of using proprietary data, we compared a set of 35 million publicly shared check-ins with a set of over 400 million private query logs recorded by a commercial hotel search engine. Although generated by users with fundamentally different intentions, we find common conclusions may be drawn from both data sources, indicating the viability of publicly shared location information to complement (and replace, in some cases), privately held location information.

Second, we introduced a couple of prototypes of new geo-social information systems that utilize the collective intelligence from the emerging geo-social footprints. Concretely, we proposed an activity-driven search system, and a local expert finding system that both take advantage of the collective intelligence. Specifically, we studied location-based activity patterns revealed through location sharing services and found that these activity patterns can identify semantically related locations, and help with both unsupervised location clustering, and supervised location categorization with a high confidence. Based on these results, we showed how activity-driven semantic organization of locations may be naturally incorporated into location-based web search. In addition, we proposed a local expert finding system that identify top local experts for a topic in a location. Concretely, the system utilized semantic labels that people label each other, people's locations in current location-based social networks, and can identify top local experts with a high precision. We also observe that the proposed local authority metrics that utilize collective intelligence from expert candidates' core audience (list labelers), significantly improve the performance of local experts finding than the more intuitive way that only considers candidates' locations.

7.2 Future Directions

In terms of the future work, we are quite interested in the following directions:

- **Modeling Group-based Mobility Patterns**: In this dissertation, we analyzed a set of 22 million check-ins from location sharing services, and modeled individual user' mobility pattern using her check-ins. Recently, Cho et al. [21] studied check-ins from Gowalla and Brightkite, and observe that users' mobility patterns are constrained by both geographical and social factors. Specifically, they observe that social relationships can explain about 10% to 30% of all human movement, while periodic behavior explains 50% to 70%. A natural follow up for both ours and Cho's work is to study group-based human mobility patterns (e.g., flock behavior). Interesting questions to study include: when and where do people check-in with friends together more frequently? Do users of location sharing services still follow reproducible patterns when they check in with friends together? What are the differences between the group-based mobility patterns and individual mobility patterns observed in location sharing services?

- **Location Privacy**: In the introduction, we discussed that one of the challenges for better understanding and leveraging emerged geo-social footprints is location privacy. Though not studied in this dissertation, location privacy still remains a quite interesting problem that we are interested to explore. Specifically, we are interested in three aspects that concern location privacy: (i) learning the usage of location sharing feature adopted in current location-based social networks; (ii) analyzing what is the percentage of users that revealed their fine granular home locations; and (iii) studying spatio-temporal anomalies (could potentially be spam or malicious manipulations of geo-tagged content) in location-based social networks.

- **Real-Time Crowd-Powered Geo-Social System**: As the next step for realizing geo-social information system, we are interested to study how to "close the loop" between "experts" in the location-based crowds and high-value stakeholders who wish to interact with them. The key motivating idea is to link crowdsourcing approaches popularized for human computation (e.g., Amazon Mechanical Turk, the ESP Game, Soylent, and others [10, 11, 31, 41, 61, 80, 91, 116, 131]) to the real-time location-based crowds that manifest in the wild, for building in situ crowd-powered geo-social systems. For example, successfully connecting an earthquake-related crowd to recovery experts as the disaster unfolds could dramatically improve resource allocation - To which areas should emergency responders be sent? Where should unmanned aerial vehicles focus their data collection? – during the critical, developing moments after an emergency when information can have the greatest impact. Beyond emergency management, many domains can benefit from access to and engagement with location-based crowds, including epidemiological and disease control experts searching for evidence of new outbreaks and the reaction of the public to new vaccines and municipalities interested in responded to local events (like the recent Vancouver riots).

  Specifically, the key challenges of the real-time crowd-powered geo-social system include: (i) who can issue tasks; (ii) crowd matching for tasks; and (iii) incentivizing crowds. In this dissertation research, we tackled the second challenge by introducing the collective-intelligence powered local expert finding system, which identify potential experts to target according to both perspectives of content and space. However, there is also a perspective of time that is missing for appropriate crowd matching. In our future work, we are interested in extend our current research on identifying top local experts for a particular topic (task) considering all the three perspectives of content, space, and time. In addition, we are also interested to explore the territory of the other two challenges to study the research questions of who can issue tasks, and how to best incentivize experts in the crowds.

REFERENCES

[1] Inc. Advameg. Profiles of all u.s. cities, 2008. http://www.city-data.com.

[2] Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffer. Web-a-where: geotagging web content. In *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval*, pages 273–280. ACM, 2004.

[3] Judd Antin, Marco de Sa, and Elizabeth F Churchill. Local experts and online review sites. In *Proceedings of the ACM 2012 conference on computer supported cooperative work companion*, pages 55–58. ACM, 2012.

[4] Saeid Asadi, Chung-Yi Chang, Xiaofang Zhou, and Joachim Diederich. Searching the world wide web for local services and facilities: A review on the patterns of location-based queries. In *Advances in web-wge information management: 6th international conference*, WAIM '05, pages 91–101. Springer, 2005.

[5] Kevin Atkinson. Kevin's word list, 2007. http://wordlist.sourceforge.net.

[6] Lars Backstrom, Jon Kleinberg, Ravi Kumar, and Jasmine Novak. Spatial variation in search engine queries. In *Proceedings of the 17th international conference on World Wide Web*, pages 357–366. ACM, 2008.

[7] 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.

[8] Krisztian Balog, Leif Azzopardi, and Maarten De Rijke. Formal models for expert finding in enterprise corpora. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval*, pages 43–50. ACM, 2006.

[9] Alastair R. Beresford and Frank Stajano. Location privacy in pervasive computing. *Pervasive Computing, IEEE*, 2(1):46–55, 2003.

[10] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on user interface software and technology*, UIST '10, pages 313–322, New York, NY, USA, 2010. ACM.

[11] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and Tom Yeh. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on user interface software and technology*, UIST '10, pages 333–342, New York, NY, USA, 2010. ACM.

[12] Thorsten Brants. Inter-annotator agreement for a german newspaper corpus. In *Proceeding of the 2nd international conference on language resources and evaluation*, 2000.

[13] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.

[14] US Census Bureau. Census 2000 u.s. gazetteer, 2002. http://www.census.gov/geo/www/gazetteer/places2k.html.

[15] Orkut Buyukkokten, Junghoo Cho, Hector Garcia-Molina, Luis Gravano, and Narayanan Shivakumar. Exploiting geographical location information of web pages. In *Proceedings of the ACM SIGMOD workshop on the web and databases*, 1999.

[16] Christopher S Campbell, Paul P Maglio, Alex Cozzi, and Byron Dom. Expertise identification using email communications. In *Proceedings of the twelfth international conference on information and knowledge management*, pages 528–531. ACM, 2003.

[17] 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. ACM, 2010.

[18] Zhiyuan Cheng, James Caverlee, Kyumin Lee, and Daniel Z. Sui. Exploring millions of footprints in location sharing services. In *Proceedings of the fifth international conference on weblogs and social media*. AAAI, 2011.

[19] Ed H. Chi. Who knows?: searching for expertise on the social web: technical perspective. *Commun. ACM*, 55(4):110–110, April 2012.

[20] Steve Chien and Nicole Immorlica. Semantic similarity between search engine queries using temporal correlation. In *Proceedings of the 14th international conference on World Wide Web*, pages 2–11. ACM, 2005.

[21] 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. ACM, 2011.

[22] Karen Church and Nuria Oliver. Understanding mobile web and mobile search use in today's dynamic mobile landscape. In *Proceedings of the 13th international conference on human computer interaction with mobile devices and services*, MobileHCI '11, 2011.

[23] Massimiliano Ciaramita, Vanessa Murdock, and Vassilis Plachouras. Online learning from click data for sponsored search. In *Proceedings of the 17th international conference on World Wide Web*, pages 227–236. ACM, 2008.

[24] Teske Coletta. Seen, by mahaya: A tool to sort the mess of social media. http://www.popularmechanics.com/technology/gadgets/news/seen-by-mahaya-a-tool-to-sort-the-mess-of-social-media-15510890/, 2013. [Online; accessed Sep 24, 2013].

[25] Computing Community Consortium. From gps and virtual globes to spatial computing - 2020: The next transformative technology. Technical report, 2013.

[26] Henriette Cramer, Mattias Rost, and Lars Erik Holmquist. Performing a check-in: emerging practices, norms and'conflicts' in location-sharing using foursquare. In *Proceedings of the 13th international conference on human computer interaction with mobile devices and services*, MobileHCI '11, 2011.

[27] Jeremy W. Crampton, Mark Graham, Ate Poorthuis, Taylor Shelton, Monica Stephens, Matthew W. Wilson, and Matthew Zook. Beyond the geotag: situating big data and leveraging the potential of the geoweb. *Cartography and Geographic Information Science*, 40(2):130–139, 2013.

[28] David J. Crandall, Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg. Mapping the world's photos. In *Proceedings of the 18th international conference on World Wide Web*, pages 761–770. ACM, 2009.

[29] Justin Cranshaw, Raz Schwartz, Jason I. Hong, and Norman M. Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In *Proceedings of the sixth international conference on weblogs and social media*. AAAI, 2012.

[30] Clodoveu A Davis Jr, Gisele L Pappa, Diogo Rennó Rocha de Oliveira, and Filipe de L Arcanjo. Inferring the location of twitter messages based on user relationships. *Transactions in GIS*, 15(6):735–751, 2011.

[31] Jia Deng, Alexander C. Berg, Kai Li, and Li Fei-Fei. What does classifying more than 10,000 image categories tell us? In *Proceedings of the 11th european conference on computer vision: part v*, ECCV'10, pages 71–84, Berlin, Germany, 2010. Springer-Verlag.

[32] Junyan Ding, Luis Gravano, and Narayanan Shivakumar. Computing Geographical Scopes of Web Resources. In *VLDB '00: Proceedings of the 26th international conference on very large data bases*, pages 545–556, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[33] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric Xing. A latent variable m del for geographic lexical variation. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1277–1287. Association for Computational Linguistics, 2010.

[34] Mack Eric. Twitter, facebook, foursquare: Tools of the modern burglar? `http://news.cnet.com/8301-17938_105-57410056-1/twitter-facebook-foursquare-tools-of-the-modern-burglar/`, 2012. [Online; accessed Sep 18, 2013].

[35] Andrea Esuli and Fabrizio Sebastiani. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *In Proceedings of the 5th*

conference on language resources and evaluation (LREC'06), 2006.

[36] Christos Faloutsos, M. Ranganathan, and Yannis Manolopoulos. Fast subsequence matching in time-series databases. *SIGMOD Rec.*, 23(2):419–429, June 1994.

[37] Clay Fink, Christine Piatko, James Mayfield, Tim Finin, and Justin Martineau. Geolocating blogs from their textual content. In *AAAI 2009 spring symposia on social semantic Web: where Web 2.0 meets Web 3.0*, 2009.

[38] Joseph L Fleiss, Jacob Cohen, and BS Everitt. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72(5):323, 1969.

[39] Foursquare. Cheating, and claiming mayorships from your couch, 2010. http://blog.foursquare.com/2010/04/07/503822143/.

[40] Foursquare. About Foursquare. `https://foursquare.com/about`, 2013. [Online; accessed Sep 18, 2013].

[41] Michael J. Franklin, Donald Kossmann, Tim Kraska, Sukriti Ramesh, and Reynold Xin. Crowddb: answering queries with crowdsourcing. In *Proceedings of the 2011 international conference on management of data*, SIGMOD '11, pages 61–72, New York, NY, USA, 2011. ACM.

[42] Dario Freni, Carmen R. Vicente, Sergio Mascetti, Claudio Bettini, and Christian S. Jensen. Preserving location and absence privacy in geo-social networks. In *Proceedings of the 19th ACM international conference on information and knowledge management*, pages 309–318. ACM, 2010.

[43] Tak-chung Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181, 2011.

[44] Qingqing Gan, Josh Attenberg, Alexander Markowetz, and Torsten Suel. Analysis of geographic queries in a search engine log. In *Proceedings of the first international workshop on location and the web*, LOCWEB '08, pages 49–56. ACM, 2008.

[45] Debarchana (Debs) Ghosh and Rajarshi Guha. What are we tweeting about obesity? mapping tweets with topic modeling and geographic information system. *Cartography and Geographic Information Science*, 40(2):90–102, 2013.

[46] Saptarshi Ghosh, Naveen Sharma, Fabricio Benevenuto, Niloy Ganguly, and Krishna Gummadi. Cognos: crowdsourcing search for topic experts in microblogs. In *Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval*, pages 575–590. ACM, 2012.

[47] Scott A. Golder, Dennis M. Wilkinson, and Bernardo A. Huberman. Rhythms of social interaction: Messaging within a massive online network. In *Proceedings of the third communities and technologies conference*, 2007.

[48] Marta C. Gonzalez, Cesar A. Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.

[49] Taher H Haveliwala. Topic-sensitive pagerank. In *Proceedings of the 11th international conference on World Wide Web*, pages 517–526. ACM, 2002.

[50] Kingsley E. Haynes and A. Stewart Fotheringham. *Gravity and Spatial Interaction Models*, volume 2. Sage, Beverly Hills, CA, 1984.

[51] Raymond Heatherly, Murat Kantarcioglu, and Bhavani Thuraisingham. Social network classification incorporating link type values. In *IEEE*

intelligence and security informatics, pages 19–24. IEEE, 2009.

[52] Brent Hecht and Darren Gergle. On the ”localness” of user-generated content. In *Proceedings of the 2010 ACM conference on computer supported cooperative work*, 2010.

[53] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. Tweets from justin bieber’s heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 237–246. ACM, 2011.

[54] Tony Hey, Stewart Tansley, and Kristin M. Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009.

[55] Nicolas E. Humphries et al. Environmental context explains Lévy and Brownian movement patterns of marine predators. *Nature*, 465(7301):1066–1069, June 2010.

[56] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining*, pages 133–142. ACM, 2002.

[57] Panos Kalnis, Gabriel Ghinita, Kyriakos Mouratidis, and Dimitris Papadias. Preventing location-based identity inference in anonymous spatial queries. *Knowledge and Data Engineering, IEEE Transactions on*, 19(12):1719 –1733, 2007.

[58] Krishna Kamath, James Caverlee, Kyumin Lee, and Zhiyuan Cheng. Spatio-temporal dynamics of online memes: A study of geo-tagged tweets. In *Proceedings of the 22nd international conference on World Wide Web*, pages

667–678. International World Wide Web conferences Steering Committee, 2013.

[59] Gabriella Kazai, Natasa Milic-Frayling, Tim Haughton, Natalia Manola, Katerina Iatropoulou, Antonis Lempesis, Paolo Manghi, and Marko Mikulicic. Connecting the local and the online in information management. In *Proceedings of the 19th ACM international conference on information and knowledge management*, pages 1941–1942. ACM, 2010.

[60] Gary King. Ensuring the data-rich future of the social sciences. *Science*, 331(6018):719–721, 2011.

[61] Aniket Kittur, Boris Smus, and Robert E. Kraut. Crowdforge: Crowdsourcing complex work. In *CHI '11 human factors in computing systems*, CHI '11, New York, NY, USA, 2011. ACM.

[62] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[63] Anagha Kulkarni, Jaime Teevan, Krysta Svore, and Susan Dumais. Understanding temporal query dynamics. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 167–176. ACM, 2011.

[64] Juhi Kulshrestha, Farshad Kooti, Ashkan Nikravesh, and Krishna P. Gummadi. Geographic dissection of the twitter network. In *Proceedings of the sixth international conference on weblogs and social media.* AAAI, 2012.

[65] 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.

[66] Theodoros Lappas, Kun Liu, and Evimaria Terzi. A survey of algorithms and systems for expert location in social networks. In *Social Network Data Analytics*, pages 215–241. Springer, 2011.

[67] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Lszl Barabsi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Computational social science. *Science*, 323(5915):721–723, 2009.

[68] Kyumin Lee, James Caverlee, and Steve Webb. Uncovering social spammers: Social honeypots + machine learning. In *Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval*, pages 435–442. ACM, 2010.

[69] 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.

[70] Linna Li, Michael F. Goodchild, and Bo Xu. Spatial, temporal, and socioeconomic patterns in the use of twitter and flickr. *Cartography and Geographic Information Science*, 40(2):61–77, 2013.

[71] Rui Li, Shengjie Wang, Hongbo Deng, Rui Wang, and Kevin Chen-Chuan Chang. Towards social user profiling: unified and discriminative influence model for inferring home locations. In *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1023–1031. ACM, 2012.

[72] Warren T. Liao. Clustering of time series dataa survey. *Pattern Recognition*, 38(11):1857–1874, 2005.

[73] Jia Lin and Alexander Halavais. Mapping the blogosphere in america. In *Workshop on the weblogging ecosystem at the 13th international World Wide Web conference*, 2004.

[74] Jack Lindamood, Raymond Heatherly, Murat Kantarcioglu, and Bhavani Thuraisingham. Inferring private information using social network data. In *Proceedings of the 18th international conference on World Wide Web*, pages 1145–1146. ACM, 2009.

[75] Janne Lindqvist, Justin Cranshaw, Jason Wiese, Jason Hong, and John Zimmerman. I'm the mayor of my house: Examining why people use foursquare - a social-driven location sharing application. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2409–2418. ACM, 2011.

[76] Xiaoyong Liu, W. Bruce Croft, and Matthew Koll. Finding experts in community-based question-answering services. In *Proceedings of the 14th ACM international conference on information and knowledge management*, pages 315–316. ACM, 2005.

[77] Edward Loper and Steven Bird. NLTK: the Natural Language Toolkit. In *Proceedings of the ACL-02 workshop on effective tools and methodologies for teaching natural language processing and computational linguistics*, 2002.

[78] Jiebo Luo, Dhiraj Joshi, Jie Yu, and Andrew Gallagher. Geotagging in multimedia and computer vision - a survey. *Multimedia Tools and Applications*, 51(1):187–211, 2011.

[79] Eric Malmi, Trinh Minh Tri Do, and Daniel Gatica-Perez. From foursquare to my square: Learning check-in behavior from multiple sources. In

Proceedings of the seventh international conference on weblogs and social media. AAAI, 2013.

[80] Adam Marcus, Eugene Wu, David R. Karger, Samuel Madden, and Robert C. Miller. Crowdsourced databases: Query processing with people. In *Proceedings of the 5th Biennial conference on Innovative Data Systems Research*, 2011.

[81] Alexander Markowetz, Yen-Yu Chen, Torsten Suel, Xiaohui Long, and Bernhard Seeger. Design and implementation of a geographic search engine. In *Proceedings of the ACM SIGMOD workshop on the web and databases (WebDB '2005)*, 2005.

[82] Cathy Marshall and Frank Shipman. Experiences surveying the crowd: Reflections on methods, participation, and reliability. In *ACM web science*, 2013.

[83] Hurst Matthew, Siegler Matthew, and Glance Natalie. On estimating the geographic distribution of social media. In *Proceedings of the first international conference on weblogs and social media*. AAAI, 2007.

[84] Kevin S. McCurley. Geospatial mapping and navigation of the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 221–229. ACM, 2001.

[85] Jeffrey McGee, James Caverlee, and Zhiyuan Cheng. A geographic study of tie strength in social media. In *Proceedings of the 20th ACM international conference on information and knowledge management*, pages 2333–2336. ACM, 2011.

[86] Mor Naaman, Yee Jiun Song, Andreas Paepcke, and Hector Garcia-Molina. Automatic organization for digital photographs with geographic coordinates. In *Proceedings of the 2004 joint ACM/IEEE conference on digital libraries*, pages 53–62. IEEE, 2004.

[87] Anastasios Noulas, Salvatore Scellato, Neal Lathia, and Cecilia Mascolo. Mining user mobility features for next place prediction in location-based services. In *Data mining (ICDM), 2012 IEEE 12th international conference on*, pages 1038–1043. IEEE, 2012.

[88] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. An empirical study of geographic user activity patterns in foursquare. In *Proceedings of the fifth international conference on weblogs and social media*. AAAI, 2011.

[89] Stefanie Nowak and Stefan Rüger. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, 2010.

[90] Aditya Pal and Scott Counts. Identifying topical authorities in microblogs. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 45–54. ACM, 2011.

[91] Aditya Parameswaran, Anish Das Sarma, Hector Garcia-Molina, Neoklis Polyzotis, and Jennifer Widom. Human-assisted graph search: it's okay to ask questions. *Proceedings of the VLDB Endowment*, 4(5):267–278, February 2011.

[92] Kitano Patrick and Boer Kevin. The local business owner's guide to twitter, 2009. http://domusconsultinggroup.com/wp-

content/uploads/2009/06/090624-twitter-ebook.pdf.

[93] William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. *Numerical Recipes in C: The Art of Scientific Computing*. 1986.

[94] Daniele Quercia, Neal Lathia, Francesco Calabrese, Giusy Di Lorenzo, and Jon Crowcroft. Recommending social events from mobile phone location data. In *Data mining (ICDM), 2010 IEEE 10th international conference on*, pages 971–976. IEEE, 2010.

[95] Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. A word at a time: Computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World Wide Web*, pages 337–346. ACM, 2011.

[96] Injong Rhee, Minsu Shin, Seongik Hong, Kyunghan Lee, Seong Joon Kim, and Song Chong. On the Levy-Walk nature of human mobility. *IEEE/ACM Transactions on Networking (TON)*, 19(3):630–643, 2011.

[97] Miguel Rios. The geography of tweets. `https://blog.twitter.com/2013/geography-tweets-3`, 2013. [Online; accessed Sep 18, 2013].

[98] Adam Sadilek, Henry Kautz, and Jeffery Bigham. Finding your friends and following them to where you are. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 723–732. ACM, 2012.

[99] Ryosuke Saga, Yoshihiro Hayashi, and Hiroshi Tsuji. Hotel recommender system based on user's preference transition. In *Systems, man and cybernetics, 2008. SMC 2008. IEEE international conference on*, oct. 2008.

[100] 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. ACM, 2010.

[101] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. In *Information Processing and Management*, volume 24, pages 513–523. Elsevier, 1988.

[102] Mark Sanderson and Janet Kohler. Analyzing geographic queries. In *SIGIR workshop on geographic information retrieval*, volume 2, 2004.

[103] Salvatore Scellato, Cecilia Mascolo, Mirco Musolesi, and Vito Latora. Distance matters: Geo-social metrics for online social networks. In *Proceedings of the 3rd conference on online social networks*, 2010.

[104] Salvatore Scellato, Anastasios Noulas, Renaud Lambiotte, and Cecilia Mascolo. Socio-spatial properties of online location-based social networks. In *Proceedings of the fifth international conference on weblogs and social media.* AAAI, 2011.

[105] Pavel Serdyukov, Vanessa Murdock, and Roelof van Zwol. Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval*, pages 484–491. ACM, 2009.

[106] Rohan Seth, Michele Covell, Deepak Ravichandran, D. Sivakumar, and Shumeet Baluja. A tale of two (similar) cities: Inferring city similarity through geo-spatial query log analysis. In *Proceedings of the international conference on Knowledge Discovery and Information Retrieval*, 2011.

[107] Blake Shaw. Machine learning with large networks of people and places, 2012. http://engineering.foursquare.com/2012/03/23/machine-learning-with-large-networks-of-people-and-places/.

[108] Blake Shaw, Jon Shea, Siddhartha Sinha, and Andrew Hogue. Learning to rank for spatiotemporal search. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 717–726. ACM, 2013.

[109] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-Laszlo Barabasi. Limits of Predictability in Human Mobility. *Science*, 327(5968):1018–1021, February 2010.

[110] Anthony Stefanidis, Amy Cotnoir, Arie Croitoru, Andrew Crooks, Matthew Rice, and Jacek Radzikowski. Demarcating new boundaries: mapping virtual polycentric communities through social media content. *Cartography and Geographic Information Science*, 40(2):116–129, 2013.

[111] Daniel Z. Sui. Opportunities and impediments of open gis. *Transactions in GIS*, 18(1):1–24, 2014.

[112] Yuri Takhteyev, Anatoliy Gruzd, and Barry Wellman. Geography of twitter networks. *Social Networks*, 34(1):73 – 81, 2012.

[113] Waldo R. Tobler. A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46:pp. 234–240, 1970.

[114] Twitter. Twitter's open api, 2007. http://apiwiki.twitter.com.

[115] Olga Uryupina. Semi-supervised learning of geographical gazetteers from the internet. In *Proceedings of the HLT-NAACL Workshop on the Analysis of Geographic References*, 2003.

[116] Luis von Ahn and Laura Dabbish. Designing games with a purpose. *Commun. ACM*, 51:58–67, August 2008.

[117] Chuang Wang, Xing Xie, Lee Wang, Yansheng Lu, and Wei-Ying Ma. Detecting geographic locations from web resources. In *GIR '05: Proceedings of the 2005 workshop on geographic information retrieval*, 2005.

[118] Carolyn Watters and Ghada Amoudi. GeoSearcher: location-based ranking of search engine results. *Journal of the American Society for Information Science and Technology (JASIST)*, 54(2):140–151, 2003.

[119] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270. ACM, 2010.

[120] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)*. Morgan Kaufmann, June 2005.

[121] Xiaoping Xi, Eamonn Keogh, Christian Shelton, and Li Wei. Fast time series classification using numerosity reduction. In *Proceedings of the 23rd international conference on Machine learning*, pages 1033–1040. ACM, 2006.

[122] Chen Xu, David W. Wong, and Chaowei Yang. Evaluating the geographical awareness of individuals: an exploratory analysis of twitter data. *Cartography and Geographic Information Science*, 40(2):103–115, 2013.

[123] Gui-Rong Xue, Hua-Jun Zeng, Zheng Chen, Yong Yu, Wei-Ying Ma, WenSi Xi, and WeiGuo Fan. Optimizing web search using web click-through data. In *Proceedings of the thirteenth ACM international conference on information and knowledge management*, pages 118–126. ACM, 2004.

[124] Yusuke Yamamoto. Twitter4j open-source library, 2007. http://yusuke.homeip.net/twitter4j/en/index.html.

[125] Jaewon Yang and Jure Leskovec. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 177–186. ACM, 2011.

[126] Sarita Yardi and Danah Boyd. Tweeting from the town square: Measuring geographic local networks. In *Proceedings of the Fourth international conference on weblogs and social media*. AAAI, 2010.

[127] Mao Ye, Dong Shou, Wang-Chien Lee, Peifeng Yin, and Krzysztof Janowicz. On the semantic annotation of places in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 520–528. ACM, 2011.

[128] 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*, 2010.

[129] 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.

[130] Mingxuan Yuan, Lei Chen, and Philip S Yu. Personalized privacy protection in social networks. *Proceedings of the VLDB Endowment*, 4(2):141–150, 2010.

[131] Haoqi Zhang, Eric Horvitz, Rob C. Miller, and David C.Parkes. Crowdsourcing general computation. In *CHI workshop on crowdsourcing and human computation*, 2011.

[132] Jing Zhang, Jie Tang, and Juanzi Li. Expert finding in a social network. In *Advances in Databases: Concepts, Systems and Applications*, volume 4443 of *Lecture Notes in Computer Science*. Springer, 2007.

[133] Jun Zhang, Mark S. Ackerman, and Lada Adamic. Expertise networks in online communities: structure and algorithms. In *Proceedings of the 16th international conference on World Wide Web*, pages 221–230. ACM, 2007.

[134] Vincent W Zheng, Yu Zheng, Xing Xie, and Qiang Yang. Collaborative location and activity recommendations with gps history data. In *Proceedings of the 19th international conference on World Wide Web*, pages 1029–1038. ACM, 2010.

[135] Yu Zheng, Lizhu Zhang, Zhengxin Ma, Xing Xie, and Wei-Ying Ma. Recommending friends and locations based on individual location history. *ACM Transactions on the Web (TWEB)*, 5(1):5, 2011.

[136] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th international conference on World Wide Web*, pages 791–800. ACM, 2009.

[137] Kathryn Zickuhr. Location-based services. *Pew Research Center*, September 2013.

[138] Wenbo Zong, Dan Wu, Aixin Sun, Ee-Peng Lim, and Dion Hoe-Lian Goh. On assigning place names to geography related web pages. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pages 354–362. ACM, 2005.