Title
A Framework for Analyzing Human Mobility in Online Social Networks

Author
Solomon, Zev Israel

Publication Date
2012

Peer reviewed|Thesis/dissertation

University of California

Los Angeles

A Framework for Analyzing Human Mobility in Online Social Networks

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Computer Science

by

Zev Israel Solomon

2012

Abstract of the Thesis

A Framework for Analyzing Human Mobility in Online Social Networks

by

Zev Israel Solomon

Master of Science in Computer Science

University of California, Los Angeles, 2012

Professor Mario Gerla, Co-chair

Professor Giovanni Pau, Co-chair

Social networks are becoming one of the most popular forms of communication between individuals worldwide. As more people start to use social networks and post more status updates, more information about the personal lives of individuals begins to leak out into cyberspace. In this thesis, we leverage the power of social networks and their Application Programming Interface (API) to data mine social networks.

Many social networks are beginning to add geo location information to status updates to show where users post updates from. Using the geo location information we gathered from Twitter and Foursquare, we are able to analyze the spatial and temporal patterns of individuals. Using studied heuristics, we are able to predict the activity of these individuals. These predictions are perfect for recommendation engines, law enforcement applications, and to track the spread of disease and exposure of harmful pollutants in the atmosphere.

We use the spatial and temporal information from the social networks in combination with UCLAs Vehicular Sensor Network. We seek to answer the questions of where people go and how long do they stay in one location. The Vehicular Test-Bed currently deployed in Macao, China tracks the concentration levels of carbon dioxide in the Macao region continuously. Using the social network information, we are able to estimate the exposure of individuals to these harmful pollutants.

The thesis of Zev Israel Solomon is approved.

Demetri Terzopoulos

Joseph Distefano III

Giovanni Pau, Committee Co-chair

Mario Gerla, Committee Co-chair

University of California, Los Angeles

2012

To my Father and Mother

who have supported me and motivated me

to achieve my highest goals.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank Dr. Giovanni Pau for all the hard work and dedication he put into helping me with this research. His guidance, wisdom, and knowledge were invaluable. I am very greatful to my advisor, Dr. Mario Gerla, for his knowledge and guidance during my stay as a Master's student at UCLA.

I gratefully acknowledge Dr. Rita Tse and the students at the Joint Research Lab in Ubiquitous Computing for hosting me and the Macao Politechnic Institute and the Macao Foundation for funding my travel to Macao. The knowledge and data gained during my stay in Macao was excellent.

CHAPTER 1

Introduction

The concept of Online Social Networks has been around since the early days of the Internet. The earliest social networks came in the form of news groups and bulletin board systems. One of the most popular services at the time was Usenet [HH97]. Usenet was a distributed bulletin board system that was created at Drake University. Although many may not consider it a social network like the Facebooks and Twitters of today, but they shared many commonalities, most important of which was allowing communication between users in a virtually free and un-moderated environment. As the Internet grew towards the end of the 20th century, many new social networking services and concepts began to arise. Services such as America Online and Geocities allowed users to share personal information in the form of easy to use webpages and communication mediums.

In 2002 Friendster was launched, heralding a new age of online social networking. A year later, MySpace was launched. MySpace rose to popularity and was the dominant force in the social networking space for years [Gil11]. However, the popularity of MySpace would be short lived. In 2004, Facebook launched and today, is still the most popular online social network according to comScore [com]. While Facebook dominates the market, there are many other popular online social networking websites in use today such as Twitter and Foursquare. Each bringing

their own unique spin and feature set to the social networking market.

These online social networking websites have grown in popularity over the years. Millions of users post status updates and share personal information on these websites. As these sites grow in popularity, so does the market for applications to take advantage of the popularity of these sites. Most social networking websites provide developers with an Application Programming Interface (API) to interact with and gather information from the social networking services. Today, entire companies are formed around using information from and monetizing the users of these online social networking websites. With so many updates posted to these websites, there is much information and knowledge to be gained about human social interaction.

To gain more users and maintain their current user base, online social networking websites must constantly revise and add new features to their services to keep the content fresh and exciting for the users. One of the popular features being added to these services is the idea of attaching geo-location information to status updates. Foursquare was one of the original websites to attach such geo-location information to status updates. In fact, the Foursquare service was built around the feature. Foursquare users submit status updates in the form of check-ins to venues in the area. Users say that they are at a specific venue and these check-ins are posted to their friends. Facebook launched a similar service allowing users to check-in to locations and post the information to their wall. Likewise, Twitter users can attach their location in the form of geo-tweets which show where they use the service from. As smart phones become more available and prevalent in society, it is easier than ever to attach this location. All smart phones carry a

GPS to determine location information and most social networks provide a mobile version of their service to use on these smart phone devices.

As more people use online networking services, especially those with geo-location features, much can be learned about human activity through these users. Specifically, we seek to answer the following two questions: where do people go and how long do they stay in a single location. Being able to answer these questions for a given user of these online social networking websites opens the door to many important applications to tie in to these networks. Applications such as recommendation engines, law enforcement, medical, and commercial applications can benefit from knowing where somebody goes and how long they stay in a location. By analyzing the multitude of status updates and changes in locations of individuals, we are able to determine popular locations and length of stay at these locations for many users.

Recommendation engines can be built around knowing where people go. With this information, a recommendation engine can target advertisements and deals to individuals who have a high probability of being at a specified location. Many small businesses can also benefit from such a system by targeting and pulling in users from nearby popular locations and hot spots. Businesses can also target discounts and special offers to specific times of day when they know that business will be slow. Such recommendation engines can become far more accurate and powerful than those in service today. This is due to the fact that a recommendation engine built on social network data works with real data from individuals in the area.

Law enforcement can also benefit from such an application in tracking suspects

in a criminal case. If a known suspect is also a user of online social networking websites with geo location enabled services, law enforcement can leverage the power of our system to determine, with a high probability, the location and trends of such a suspect. Such a system can lead to reduced crime and aid in the recapture of fugitives living at large.

There are many medical applications of geo-location information in online social networks. Most notably, medical staff can use the information on these online social networks to track the spread of disease and pandemics across the globe. As disease spreads from person to person, people begin to post status updates related to the disease. By following these updates, it is possible to determine where and how fast such a disease spreads. This technology can also be used to reach out to underdeveloped and underrepresented communities which may not have access to the best medical care. Hospitals can use this information in preventative screening and testing to help reduce leading preventable causes of death in these communities.

Pollution and exposure to harmful pollutants is a very important field of study and research today. In the city of Macao in China, the University of California, Los Angeles and the Macao Polytechnic Institute have deployed a Vehicular Sensor Network [LMG09] aimed at monitoring and mapping pollution in Macao. The vehicles are equipped with a multitude of sensors as they drive around the city streets to collect and gather data on pollution in the city. This data is later uploaded and mapped to show a constantly moving pollution map. By leveraging the power of online social networking websites, we are able to measure the exposure of these individuals to these harmful pollutants. If we can determine how long an

individual stays in a location, we can correlate this data with the data provided by the vehicular sensor. This allows us to, in real time, calculate the exposure of individuals to these pollutants. Such a solution can be employed by many municipalities and medical facilities to improve the quality of life of individuals in pollution rich environments.

Such a rich source of information available online is not without privacy concerns. As we post more about our personal lives to these online social networking websites, we begin to give up more of our privacy. There is a serious concern about who has access to our status updates and who many exploit them for malicious means. However, most of these concerns are preventable and are caused by the users themselves. All social networking sites provide privacy settings that the users can set for themselves. These privacy options prevent their information from being posted outside of their own contacts. Most users, however, do not know about or do not choose to configure the privacy options on the web sites. Due to this, most of the data on these online social networks are, in fact, public. Our data mining software is designed to be fully privacy compliant. We only use those API calls that are public and we only access the public data on these websites. Although this limits some of the data that we are able to collect, there is still plenty of information available from those users who retain the public nature of their profile information.

Social networking websites are becoming increasingly prevalent in our lives today. As the Internet spreads to more and more individuals worldwide, the user base of these online social networks is constantly growing larger by the second. Many of these online social networks have features that allow users to attach

geo-location information to their status updates. By leveraging the geo-location information from millions of status updates, we seek to answer the questions of where users of these social networking websites are going and how long they stay in a single location. The answers of these questions have many important applications for many different industries. We seek to use this information to determine the exposure to harmful pollutants in the city of Macao in China.

CHAPTER 2

Related Work

Social Networking is an emerging field in Computer Science. We study the characteristics of the Social Network user base in order to gain a better understanding of the types of users that use these applications. Next, we study the connectivity between users and geo-spatial clustering of users. We then examine applications and case studies where geo-spatial properties of status updates in Social Networks are used to solve complex problems. Finally, we address privacy concerns that arise with the increasing popularity and prevalence of Online Social Networks.

2.1 User Classification

In order to determine that our methodology applies to the general population, we must first determine that the users of Online Social Networks are not simply a small, biased subset of the population. We must prove that these users are a diverse, representative set of users. Blake Shaw, Data Scientist at Foursquare found that at every second, 35 people check-in to a location on Foursquare [Sha]. As of October 2011, Foursquare had data on over 25 million places across the globe. Such a large data set provides a rick source of information about human behavior. Machine learning algorithms are then applied to classify and discern patterns of

human behavior on Foursquare. Venkatraman found in 2010 that Americans spend over 25 percent of their time using Online Social Networks [S10]. His study found that while Google is still the dominant force in web search, Facebook has become the most visited website. His study also shows that since social networking is a very new field, scholarly research on the impact of social networks is lacking.

There are also several psychology studies that aim to understand the state of mind behind the users of online social networks. Ellison et al. conducted a study that examines the actual audience of Facebook [ESL07]. The study looks at users and how they relate to social capital. Social capital as defined by Bourdieu and Wacquant is the sum of the resources, actual or virtual, that accrue to an individual or a group by virtue of possessing a durable network of more or less institutionalized relationships of mutual acquaintance and recognition [BW92]. Individuals use this social capital to gain information, form personal relationships, or organize groups. The study found that Facebook users interacted with all three types of social capital and that usage provides great benefits to those with low self-esteem.

The social network user base has many similarities to the video game user base in terms of casual and hardcore users. Kuittinen et al. presents a study on the casual and hardcore usage of video games [KKN07]. They find that accessibility of the games determines the user base. Games such as free to play or micro-transaction games were found to be casual. However, casual only refers to the type of game. Social networks would fall under the casual category as well due to the accessibility of them. Joinson et al. presents a similar study which looks at the motivation behind users of online social networks [Joi08]. The study found

that users have a variety of uses and gratification that arise from social networks. These gratifications lead to patterns of usage which cause increased frequency of use and time spent on the site. The study also shows that many users feel that the default privacy setting is too restrictive to allow them to discover new people. There are some privacy concerns, however, that arise form usage of online social networking sites, however, users find that the need to discover new people outweighs the need to maintain their privacy. Social networking sites appeal to a very large audience and are very accessible. Usage of online social networks has been found to be a very fulfilling and gratifying experience. Since social networks greatly fuel social capital, users tend to spend more hours and more frequently visit these networks. Users then tend to share more information about themselves to others.

2.2 Friend Connectivity

Since we established that the users of Online Social Networks are representative of the general population, we must now examine how users are connected to one another. Social networks are organized in such a way that people share information with their friends. By analyzing the connectivity between users, we can see whether or not these relationships follow similar trends as those in real life.

In a study performed by Catanese et al [CDF11], a Facebook crawler was constructed to try and analyze the social structure of Facebook. They use the crawler to take a sample of millions of connections between Facebook users. Organized as

a graph, they came to the conclusion that Social Networks are highly connected. They also found that a strongly connected component exists in the social graph which shows that there are some users with very high degrees of connections with other users. These highly connected users are of particular interest to data collection applications because of the amount of data that can be discovered from these users. Another study performed by Scellato et al. [SMM10] analyzes four more online social networking websites. The study reinforces the Catanese study by showing that these social networks have similar structures including a highly connected component.

Scellato performs another study to look at the socio-spatial properties of users by examining the locations of these users [Sce11]. They look at the distance between users who have connections. In the study, they found that regardless of the service they use, users exhibit friendship connections that exhibit similar properties to physical connections. In 2010, Pete Warden gathered data from over 210 million public Facebook profiles [Warb]. He developed a web crawler to gather data from the public pages of Facebook users. Using the data, Warden was able to discover patterns in user behavior. The data showed that users tend to form clusters depending on users geographic location in the United States. These clusters also have similar trends and popular topics. The study by Scellato and Wardens crawler show that online social networks are heavily influenced by trends in a users physical surroundings.

2.3 Analyzing Trends and Patterns in Social Networks

The volume of user data posted to online social networks every day is incredible. Given access to the data, applying statistical models to social network data can give us insight into the daily lives of individuals who use these social networking websites. Vicente et al. discusses the proliferation of Geo-social networks [RFB11]. In their study, they discuss context-aware services that combine location and user content. There are many different types of Geo-social networking websites, but most of them share a commonality in that each user shares their physical location of where they sent their status updates. Vicente discusses how users use these venues to capture real-life relations or indicate affinities or common interests. Most networks such as Foursquare are location-centric in that users can retrieve content by going to a locations page. Twitter, on the other hand, is user-centric where geo-coded tweets are tied to individual users.

One of the most obvious applications for geo-social networking data is that of a recommendation engine. Ye et al. presents a recommendation engine that exploits social and geographical characteristics of online social networking websites [YYL10]. They develop a method called geo-measured friend-based collaborative filtering that is based on ratings provided by friends. The social network they aim to analyze is Foursquare. They describe difficulties in gathering data due to the privacy settings on Foursquare.

A study by Cho et al. looks at analyzing human mobility patterns in online social networks [CML11]. They use cell phone location data as well as online social networking data in their analysis. They construct a probabilistic model

which looks at trends of check-ins over time on the popular Foursquare online social networking website. Their model shows that users tend to post updates from the same locations at the same time of day with a high probability. They show on a map where users tend to be at specific times of day, namely work and home. They use the probability model to create areas of activity to show where a user is most likely to be at a certain time of day.

Social network geo-location data has many medical applications as well. An important application of this data is in tracking the spread of disease. In a study by Lampos and Cristianini, they track the spread of the 2009 H1N1 flu pandemic in the United Kingdom. In their study they use both the geo-location information embedded in Twitter updates as well as the text itself [LC10]. By looking for specific keywords related to the flu pandemic, they are able to assign a flu score to specific tweets in the United Kingdom. By analyzing these Tweets, they are able to monitor the spread of the disease by looking at where the tweets originate from. At the end of the study, they compare their results with those of the Health Protection Agency and find over a 95% correlation. Similar studies have been done which reinforce the notion of being able to detect large magnitude events using social networks [HP09]. These studies look at similar emergency situations such as earthquake detection in Japan [SOM10]. These studies use similar methods in analyzing the popular Twitter service to infer large scale events based on Twitter activity. The studies show that online social networking websites, especially those with geo-location features are very strong and accurate sources for information related to large scale events.

2.4 Privacy Concerns

The vast data available from online social networking websites does have several privacy concerns attached. With the proliferation of data and status updates from these websites, more and more information about the personal lives of these users becomes public, including their location in the case of geo-location enabled online social networks. The study by Vicente et al. addressed some privacy concerns that they encountered in the design of their crawling application [RFB11]. There are many challenges faced with trying to maintain privacy in geo-location enabled online social networking services. A large concern is that many of these services provide the ability to tag users. This exposes the information to an even larger group of users causing the individual to have very little control over their own public data. The two major threat categories described by Vicente are the release of sensitive location information and re-identification through location. Sensitive location information refers to the specificity of the information. For instance, just revealing that you live in Los Angeles is not a very large threat, but revealing the specific restaurant or apartment building you are at is. Likewise, re-identification means that a third party can narrow down who you are by analyzing where people are going and narrow down a list of users that may be present at that location. The three notions of privacy then, as described by Vicente are location, absence, and co-location privacy. Social networks can address location privacy by revealing only more generalized location information. Likewise, absence refers to not being at a location. For instance, if you are not at home and a potential burglar knows this information, they could use that to their advantage to steal from a user. Currently, online social networking websites do not address this issue. Lastly,

co-location privacy refers to multiple users being together at a specific location. This information can be potentially damaging to relationships and other sensitive situations. Most social networks do not address this either as information is available to all friends in most cases.

Vicente proposes several techniques to preserve privacy in these networks. Many of these techniques involve poisoning the data set with false location information. Although this is a valid method of privacy protection, it can defeat the purpose of such social networking websites. A study by Acquisti and Gross shows another privacy concern with online social networks in general [GA05]. The main concern raised by the paper is that users do not change or are not aware of privacy settings that exist on the social networking websites. Almost all social networks offer extra privacy settings that users can activate to further protect their profiles. However, for many of these sites, the default setting is more public. Many privacy concerns lie with user responsibility because they are able to set these additional privacy settings but they do not or are not aware of them.

Another study by Lam et al. in Taiwan discusses a similar privacy concern [LCC08]. The study focuses on Wretch, the largest online social networking website in Taiwan. The principle problem described in the paper is that of involuntary information leakage. Since users on Wretch are able to annotate their friends profiles, it is possible to infer private information from that profile. They found that the majority of profiles on Wretch had private information exposed. Although it is the users responsibility to maintain the correct privacy settings on their profiles, due to the nature of some online social networks, it is difficult to maintain privacy as information begins to spread between large groups of users. This is especially

true for geo-location enabled online social networking websites. These websites have users post their current location which gets spread between many users in the network. There are many valid privacy concerns related to online social networks, however for the most part, many of these concerns can be addressed with more diligence in what users post and what settings are enabled or disabled in their profile settings.

CHAPTER 3

Design Principles

We designed UCLASocial as a system to gather data from various online social networking websites. We needed an application that could gather large amounts of data continuously. Since social network users can send status updates at any time of the day, the application had to be autonomous so that a contiguous stream of data could be obtained. Data from a single point in time is not useful since we intend to analyze the motion of individuals over periods of time. It is also impossible to gather complete data from these sites since the beginning of time. There are also limitations in the Application Programming Interface (API) of the various social networking websites that limit search results as well. We chose to run the application at regular intervals, indexing and storing the data server side, to avoid these limitations.

Different social networks have different API calls as well as different data structures. We needed an application robust enough to be able to handle API calls and data structures from several sites. We designed our application to be modular in order to support different social networks. This modularity gives us the ability to collect different types of data while storing the data in a structure that is useful to our analysis. Social networks provide us with an abundance of superfluous data that is not a part of analysis. By making our application

modular, we are able to focus on only the data that is of interest to us while being flexible enough to deal with a multitude of social network API calls.

Lastly, we needed an application that would allow for user authentication as well as the ability to communicate with social networks. All of the social networks we worked with have API calls that are performed by sending HTTP requests to their servers. These two properties of social network APIs lends itself well to web based solutions.

With these requirements, we chose to use the Perl Dancer web framework [Per] for the front-end of UCLASocial while using the MySQL database system [Mys] in the back-end. Perl and MySQL work particularly well together and are proven to be robust and flexible enough for our solution. Perl Dancer is a web framework that allows for a different script to run depending on the HTTP request sent to the server. These HTTP requests are then sent at regular intervals so that data can be collected in an unattended fashion.

3.1 Twitter Geo-Tweet Search Engine

Twitter provides a large quantity of data due to the public nature of the social networking site. Since by default all profiles are set to public, Twitter lends itself well to anonymous data mining. Twitter also does not require any user authentication to use their API. Since Twitter updates are geocoded with the location of the user at the time of the status update, Twitter is a perfect resource for analyzing human mobility.

Our application leverages the power of Twitters powerful search API to gather

information on specific geographic areas. Since Twitter introduced geo-coding into their status updates, the ability to search based on specific locations has been implemented into their API. The search API allows UCLASocial to gather information from specific venues or areas of interest. The advantage to this approach is that we are able to remove the vast majority of the uninteresting and noisy data that is associated with Twitter.

Unlike other social networks such as Facebook that emphasize friendship and require friendship to view information, Twitter is designed to have users, for the most part, share status updates publically. Twitter uses a system of following as opposed to friendship. Following a user does not require permission and will result in you automatically receiving all of their updates on your front page news feed. This news feed contains all the updates sent by you and all users you are following. Our application takes this idea of following, but extends it to a geographic area. The result is a newsfeed being generated for each area of interest for our application.

3.2 Foursquare Checking Scraping

Foursquare is a rather unique social networking site because the main focus of status updates are locations. Users check in to a location by physically going there and using their phones GPS to send a status update. Foursquare is unique in the sense that unlike Facebook and Twitter, Foursquare requires users to attach their physical location to their updates. This requirement of attaching a physical location to check ins makes Foursquare a very interesting social network to data

mine. Using Foursquare check in data from a user, it is easily possible to trace a general path that the user took over time. This ability to trace user activity is very useful for creating location aware applications that work together with Foursquare.

The requirement to attach location to Foursquare updates leads to a multitude of privacy concerns. Due to the fact that much of the user data on Foursquare is considered sensitive, the Foursquare API relating to users is very restrictive. Each individual must authenticate themselves with the application before data collection is allowed. Data mining Foursquare on a large scale, therefore, is impractical as it would require each and every user in a given area to authenticate with the application.

This restriction on user data, however, does not apply to the Foursquare venues themselves. Venues on Foursquare are considered public and do not require permission to retrieve data. UCLASocial uses the venues feature of the Foursquare API to discover popular areas with in regions we would like to collect data from. Each venue contains information about how many people have visited the location and how many people are currently at the location. This allows UCLASocial to gain a general knowledge of how people move around through our area of interest. Over time and at different times of the day, different venues become more popular than others. This information has applications not in following users, but in recommending users where to go. A recommendation engine can use the Twitter data to determine where a user is and recommend a popular Foursquare venue nearby.

3.3 Facebook Crawling

Facebook has always been an interesting social networking site to gather data from. Users on Facebook tend to share more personal information due to the tightly knit friendship nature of the site. Before Facebook users are allowed to communicate with each other and view status updates, a mutual friendship between the users is required. Likewise, the Facebook API is just as restrictive. In order to gather any updates from a user, it is necessary to request permission. Facebook requires specific extended permissions in order to view a users news feed. Since these permissions are required of every user we need to download data from, it is very difficult to gather enough meaningful data from Facebook users.

Some Facebook users have their profiles set to public. As was discovered in [Warb] and [CDF11], Facebook users tend to form a graph structure with their friends. By traversing this friendship graph, it is possible to gather large quantities of public data from users. This data, however, is only visible on the public website and is restricted from the Facebook API. We attempted to construct a crawler that would crawl the public webpages of Facebook to explore the social graph. However, since Facebook constantly changes the layout of the web pages with every new feature added, crawling the graph and retrieving consistent and reliable results has proven very difficult. Facebook also uses large amounts of AJAX to update pages on the fly which causes our crawler to miss a lot of important data. Friends lists, for example, are generated using AJAX, so when viewing them with our crawler, we are unable to traverse the graph.

When Pete Warden crawled Facebook in 2010 [Warb], he was able to generate

a huge data set with very large quantities of user updates. He was able to reach many interesting conclusions revolving around social patterns of users in different geographic regions and also was able to view the social graph structure and related connections between users. After he made his discoveries public, he wanted to release the data set to academic institutions. Shortly after his announcement, he was sued by Facebook [Wara] and the data was subsequently disallowed from being released. Facebook later amended their developer terms of service to explicitly disallow crawling the public pages. There are many problems revolving around gathering data from public Facebook users which has prompted us to focus more on other social networks to gather data.

CHAPTER 4

UCLASocial Architecture

We designed UCLASocial as a modular and expandable solution to social network data mining. Perl Dancer provides us with the modularity we need to interact with different social networks and is also flexible enough to retrieve data in different ways and interact with our database solution. Our solution is very similar to the LAMP solution [LW02] which is very versatile and robust enough to handle real world applications. Perl Dancer is a web framework for Perl. Dancer uses a series of application scripts and templates to generate a lightweight web application. Dancer listens on a predefined port for HTTP requests. When it receives a request from a client, it will execute a specified portion of code. Dancer also allows us to respond to POST requests such as those used during authentication with Facebook and Foursquare. We chose to use Perl Dancer because it is fast, flexible, and easy to develop. We also chose Perl to leverage the power of the Comprehensive Perl Archive Network (CPAN) [CPA] which contains many libraries for social network authentication as well as statistical analysis tools which we use during data analysis.

UCLASocial follows a similar style to many social network APIs by creating various endpoints for each data gathering operation. These endpoints all follow a consistent structure by having the module name followed by the script name in the

URL. Our data gathering application consists of three modules for Foursquare, Twitter, and Facebook. The modules interact with a common database which archives the data for future data analysis. Since most sites do not provide all historical data to their API, it is important for us to maintain an archive of historical data so that we have a larger and more comprehensive data set to work with during data analysis.

4.1 Twitter Search Engine

The main source of data collection for UCLASocial is through the popular online social network Twitter. The main benefit to using Twitter for data collection is that a large portion of the status updates posted to Twitter are public. This allows for our application to forego user authentication so that we are able to simply gather data from the large public pool. Many Twitter updates now contain geographic location information of where users tweeted from. The Twitter module for UCLASocial is built gather updates from specific regions. This allows ups to gain knowledge of user interactions regionally.

Twitter provides a robust search API that allows for the fine-tuning of parameters in order to more narrowly define search. UCLASocial begins by selecting a region to search from. We originally set up the search to be on the regions of Macao, China and the UCLA campus in Los Angeles, California. The motivation behind this was to gather information in the coverage areas of the Vehicular Sensor Network being developed at UCLA [LMG09]. We later expanded this search to popular regions such as New Yorks Times Square and San Franciscos Union

Square. These venues provide us with a strong pool of users. The Twitter search API does, however, limit the number of responses to 1,500 tweets per search. Therefore, we must perform periodic search and archive operations for each region. These regional search operations provide us with data that has very coarse granularity. There are many different types of Twitter users who go to these locations. We use these coarsely grained search operations to try and classify users according to their types of Twitter usage ranging from casual to hardcore users.

Once we have selected a pool of users to examine in a region, we are then interested in their own Twitter history. Using a similar search feature on Twitter, we are able to gather a history of status updates from particular users, given that their profiles are indeed public. With these searches, we are particularly interested in user mobility. The questions we seek to answer are where people go, how long they stay at a particular location, and are there any patterns that these users tend to follow. We can download and examine the Twitter history of the users by accessing their Twitter feed. We pay particular interest to tweets which have geo-location information attached. The users we examine are those with a high concentration of geocoded tweets. Since this feature is now enabled by default for most users tweeting with smart phones, we can get a large enough sample size to begin to infer position information and patterns from these users.

Twitter gives us a good insight into user mobility. It is one of the most active online social networking websites and is also one of the most public. Since users share information publically with other users, there is a large quantity of data that can be harvested from Twitter. The search API on Twitter is especially of interest to us because it allows for very detailed and fine grained search. The

search results we receive are generally free of most noise and the vast majority of the data is useful to us. In order to properly calculate exposure to harmful pollutants from our vehicular sensor network, we need a large concentration of geo-coded data. Twitter allows us to download this data feely and publically and is an excellent resource to study human mobility on online social networks.

4.2 Foursquare Venue Check-ins

Foursquare is the ideal social network to explore for location aware applications. Foursquare was designed around using users locations as status updates. These locations are also tied to specific venues that are entered into the Foursquare network. As was shown by Shaw in [Sha], Foursquare can tell us a lot about human mobility by exploring the millions of check-ins in the Foursquare database. Unfortunately, Foursquare is not willing to share their database with us so we must develop a method to analyze check-ins using the API calls developed for Foursquare applications.

Foursquare does, however, limit the ability for applications to collect data. All applications must authenticate using the OAuth2 protocol [Oau] and must request permission from users to use the application. Due to this, it is very difficult to gain a large enough pool of users to gain meaningful data. We attempted to solicit users around campus to give permission. Although we were able to get permission from some users, we quickly found this to be a very biased pool of data. We concluded that downloading and analyzing individual check-ins would be very difficult if we want to avoid a biased set. Instead, we looked to the other

aspects of the Foursquare API and found that it would be possible to exploit the venues platform to gain some knowledge of user activity in the area.

On Foursquare, venues are considered public. Nobody outright owns the venue unlike pages on Facebook which are owned by individual people. Instead, a mayor system is put in place where the user who checks in the most becomes mayor of the venue. Since venues are considered public, the API also allows downloading of all venue data without permission. UCLASocial uses this attribute of venues in order to gather check-in data from each venue. This data gives us insight into how popular venues are at different times of day. The Foursquare API also provides us with herenow values. The herenow values try to tell us who is currently at the venue. It contains a list of specific users whose last check-in is at the current venue. This list is updated by the Foursquare servers whenever users check-in to a location. Our application polls this data at set intervals in order to determine user activity at each location. We can then see what times are more popular than others at different times of day.

Foursquare is an excellent source of information for determining user mobility. By analyzing check-in information at venues, we are able to gain knowledge of user mobility between popular venues on Foursquare. In a recommendation engine, this information tells us what locations are popular so that we can suggest locations based off popularity at a given moment in time. For the vehicular sensor network, we poll different venues within the range of the test bed. This gives us an idea of average pollution exposure in these locations. This average pollution can be used to determine if these popular locations lie within areas of high pollution. Foursquare has some of the richest mobility information available from online

social networks. Our application leverages the available public information to gain a general insight into human mobility in a particular region of interest.

4.3 Mapping User Activity

In order to determine which venues to select, we look at several features of the social networks as well as third party lists to choose locations. Twitter provides us, as party of their search API, a method to look up venues and points of interest in a given area. Foursquare has a similar feature in their explore API which provides a list of venues in a given city or geographic area. There are also several published lists which contain venues such as [Ven]. We choose popular venues to gain a large pool of users to select from. This gives us a greater opportunity to gather data from different types of users. Such locations include New Yorks Times Square, the Santa Monica Pier, and several hotels and casinos in Macao.

We use this data to gather potential users to further examine and determine activity. Once we have gathered a sufficient quantity of users from the area, we begin to create profiles for each user. These profiles contain the geographic data for each users status updates from Twitter. We can use this data to map out a users locations over time. These profiles allow us to find patterns in user movement. We can then tell whether a users location changes in a deterministic fashion. Once this is established, we can infer future location updates by using the historical data provided by the API on the social networking websites

We can use UCLASocial to produce maps of user activity over time. This allows us to find areas of high user activity at different times of day. We can

also use this information to determine whether a user uses the service enough to provide enough data for inference. Some users do not use the geo-location features enough to provide a strong enough inference as to location prediction information. We use metrics such as inter-tweet time and average distance traveled between tweets in order to classify users into categories. We also use the data from UCLASocial to determine popular venues. By analyzing the frequency of users entering these venues, we can gain information as to how many users are present at certain times of day. This allows us to determine how long they stay in a location by analyzing those tweets which occur frequently in succession at a single location. UCLASocial is a powerful data collection tool which allows us to gain high quantities of information related to human mobility in online social networking websites.

CHAPTER 5

Human Mobility Analysis of Social Network Data

The question we seek to answer in analyzing social network data is two-fold. We want to know where people are going, and how long they stay in a particular area. In order to analyze the characteristics of mobility that users exhibit, we must examine the history of status updates by the user. By looking at these updates, there is much information that can be gathered related to their activity. However, not all users provide enough data to be able to be analyzed with high accuracy. It is then necessary to observe those users who use the geo-location enabled features of online social networks enough to give us the data necessary for analysis.

5.1 User Selection

In order to successfully analyze human mobility for social network users, we must select users that use the service enough where they provide sufficient data for analysis. These users must not only use the social network service, but must also use the geo-location tagging features of the social networks as well. When selecting users from Twitter, we first gather data from popular areas around the world as

well as particular areas of interest to our study. These popular areas include New Yorks Times Square, San Franciscos Union Square, the Santa Monica Pier in Los Angeles, and other popular locations. We also gather information from Macao in China where we work with the Joint Research Lab in Ubiquitous Computing at the Macao Polytechnic Institute. We gather this data to work together with the Vehicular Sensor Network that analyzes pollution levels in Macao.

The Twitter search API allows us to gather data from particular locations with geo-location information attached. Once we selected these locations, we use the UCLASocial platform to gather data over a period of several months. The data we collect is a sample of data from several specific regions. Twitter has millions of users, many of which are either very casual users or are programs designed to send tweets. The aim of our data collection methods with UCLASocial is to identify and sub sequentially filter out the noise that is present in the Twitter feeds. Noisy data can throw off calculations and is, in general, of little to no interest for us. This noise does not help us answer the questions of where people go and how long they stay in these locations.

UCLASocial does not perform any textual interpretation or analysis of the content of the status updates. Instead, the system leverages the embedded geo-location information that is attached to Twitter updates. Although this geo-location feature is optional, we found many users that attach the data to almost every status update they send. UCLASocial gathers all tweets returned by the search API of Twitter, however, not all information contains these geo-location coordinates in the status updates. Users are able to attach a more general location to their updates, such as a city name, however we found that this information may

not be accurate or correct. Twitter also, at times, returns results for incorrect locations given search coordinates. This leads to many problems if geo-location information is not attached. Due to these concerns, and the lack of accuracy without the geo-location information, we do not consider these status updates as part of our analysis. Although this shrinks the population of users that we are able to analyze, there is still an abundance of users who attach geo-location information to the majority of their status updates on Twitter.

We have identified the types of status updates we look to analyze, so now we must determine which users we consider to be the most active users. There are many users on Twitter that only occasionally post updates to their news feeds. These users may post once a day or once a week. Since these users do not provide enough information to determine anything about their time of stay at a specific location, we consider these users to be casual users. On the other hand, there are users who post at very frequent intervals, often times in the same location. It is these users who provide us with the most information that we seek to analyze in our data set. In order to determine how to categorize users, we use several metrics on our user population. We compute the total volume of tweets as well as the number of tweets per minute of these users. We compare these values to the median of the volume and tweets per minute of our entire population. In the population of Macao, we found the following distribution of users grouped into bins in Figure 5.1. For the users in our population, we found that a relatively small group of users were responsible for the majority of tweets in an area. Furthermore, we found this trend to be consistent in other areas of interest in other cities. In the Westwood area of Los Angeles near UCLA, we find a similar distribution in

Figure 5.2.

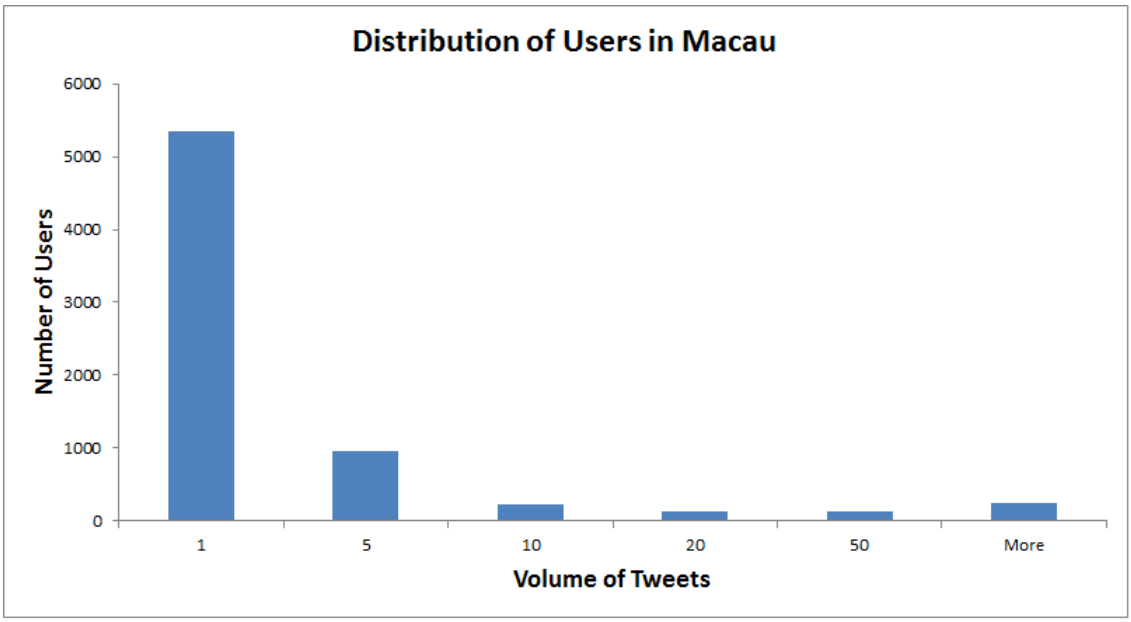

Figure 5.1: Distribution of users in Macao, grouped by number of tweets.

After finding those users with high activity, we can select and determine that these users can provide us with sufficient data to analyze their mobility in these cities. We can also remove those users who have only a single data point or too few data points. We also looked at specific venues and found similar results in terms of small numbers of users being responsible for the majority of status updates in these online social networking websites.

We conclude that these users constitute a representative sample of the total data on Twitter. Since these few users provide over half the data on the network, we can choose a smaller sub population to analyze more closely. With the smaller sub population, we can look more deeply into patterns and trends that arise out of their Twitter newsfeed updates. Since the Twitter search feature returns current Twitter updates, we have to go to these users individually and collect

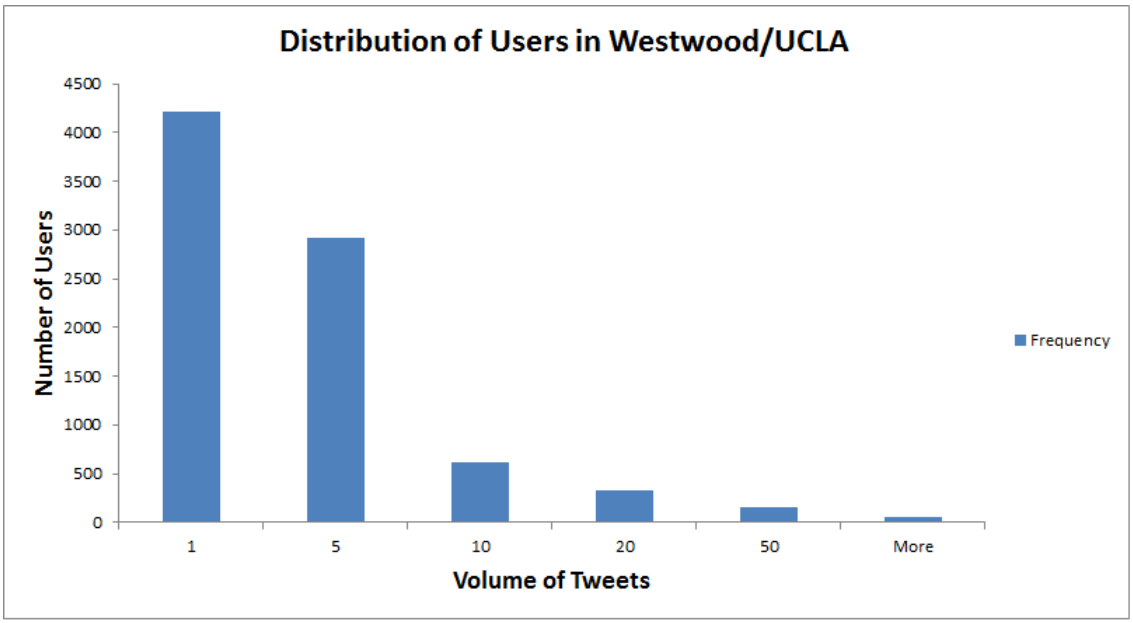

Figure 5.2: Distribution of users in UCLA, grouped by number of tweets.

their historical updates.

5.2 User Trends and Location Changes

Since we already established a smaller representative sample of our whole population, we now analyze individual users and their usage patterns. We first examine when users tend to post the most updates. We examine updates on different days of the week as well as different hours of the day. In Figure 5.3 we look at the volume of tweets on different days of the week in Macao and in Figure 5.4 we look at the different days of the week in the Westwood/UCLA area. Similarly, in Figure 5.5 we see the distribution of volume over different hours of the day. Likewise, in Figure 5.6 a similar distribution is shown in the Westwood/UCLA area.

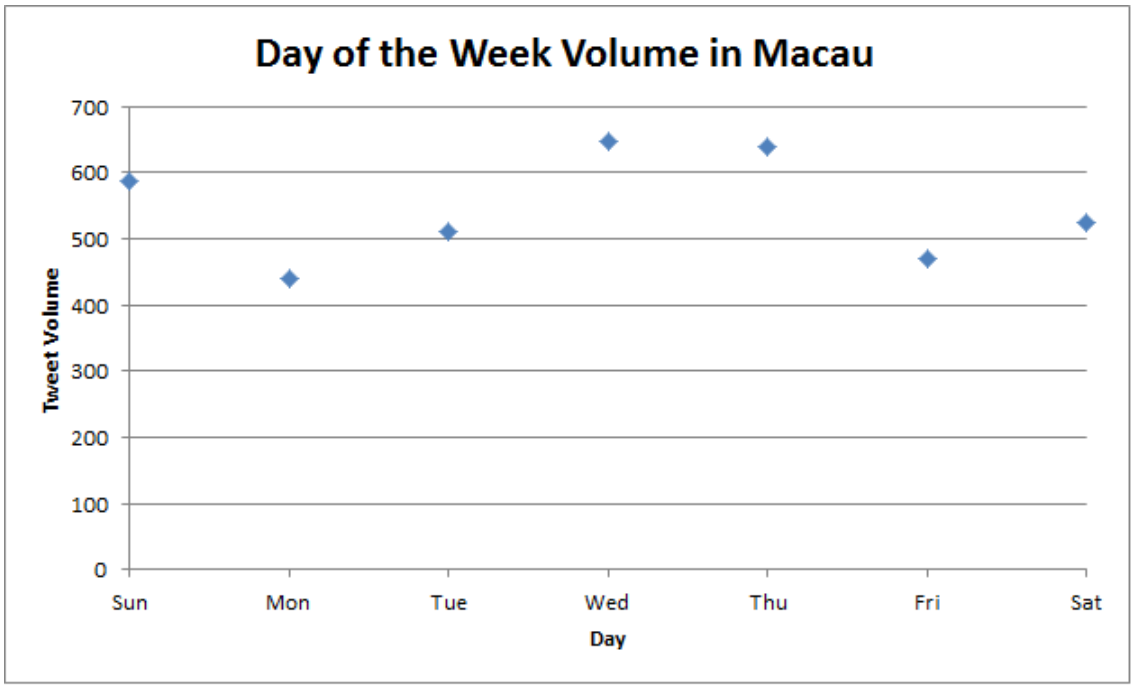

Figure 5.3: Daily volume of tweets in Macao.

We look at several metrics in determining usage patterns of social network users. We look at the time between tweets as well as the distance traveled between tweets. We can plot this on a graph for several users to see their usage patterns over time. We select these users from the representative sample we collected. From these graphs, we can further infer the type of usage that these users have. In Figure 5.7, we see the graph of a heavy user in our data set. These graphs show interesting trends with users. For the most part, we find that these users tend to post many updates from the same location. We can infer that this is their home or place of work by the frequency and time of posting. These updates tend to happen with low inter-tweet time as well as little to no change of distance. Furthermore, we find that by analyzing the multitude of different locations that a user tweets from we discover that in general, a single location of posting tends to

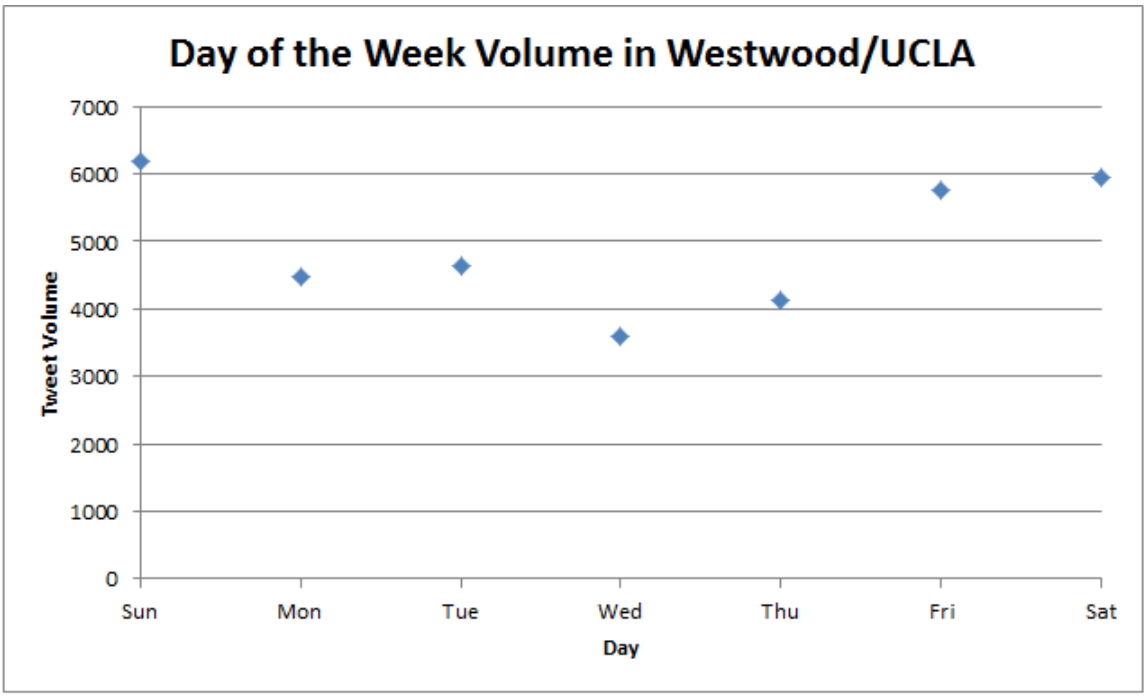

Figure 5.4: Daily volume of tweets in Westwood/UCLA.

dominate all others. This reinforces the notion that this is their home location. In Table 5.1, we examine average inter-tweet time and changes in distance between the different types of users: heavy, medium, and light. We also compare these values to those of several locations and venues of interest.

For those tweets that cause a high change in distance, similar patterns across our user base arise. We find that, in general, a large change in distance corresponds with a large inter-tweet time. This is to be expected as there is travel time that has to take place between two places that are far apart. However, when there is a large inter-tweet time, this does not necessarily correspond with a high change in distance. By further analyzing different users, we find that one of two possibilities generally occurs with these types of high inter-tweet times. Either the user traveled during that time and did not post a status update, or they stayed

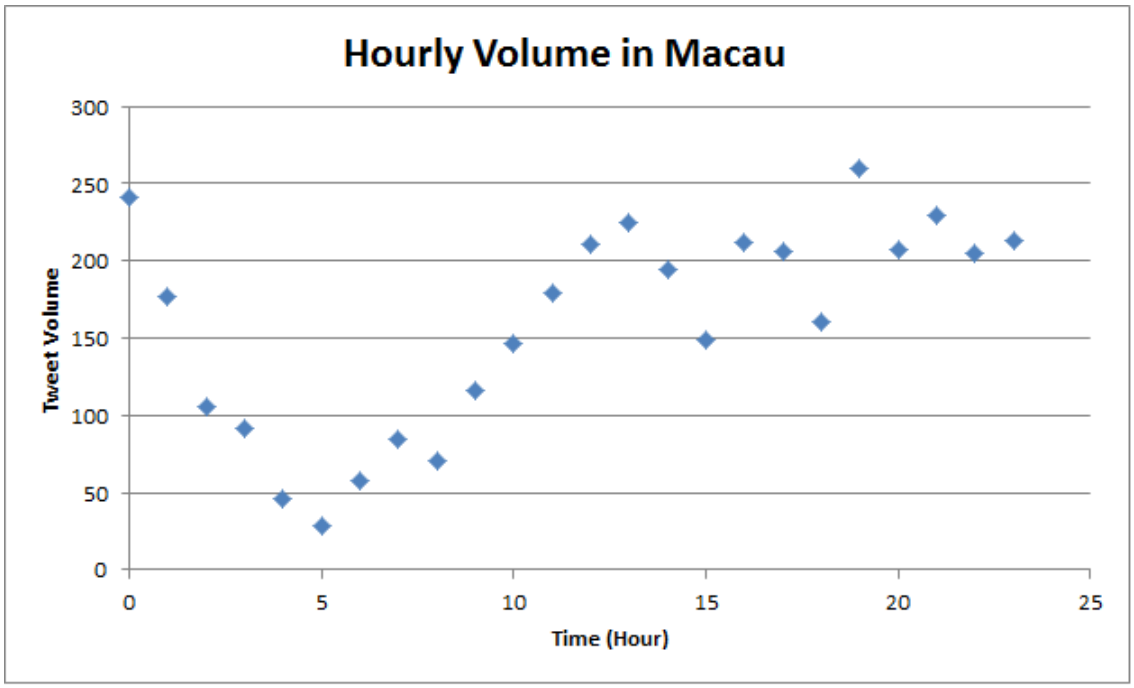

Figure 5.5: Hourly volume of tweets in Macao.

in the same location the whole time. It is difficult to determine exactly what the user did during that time since human beings tend to exhibit a high degree of variability between status updates. However, but examining their previous status updates, it is possible to determine the probability that a user traveled during this time.

When users post in quick succession with little movement, we can with a high degree of certainty, say that they stayed in that single location the whole time. Likewise, when a user posts a status update with a high inter-tweet time and a high change in distance, it can be inferred that the user was in travel during that time period. It is those status updates with high time change but little distance movement that cause a degree of uncertainty in the results. We therefore determine the certainty that a user is in a single location by analyzing those tweets

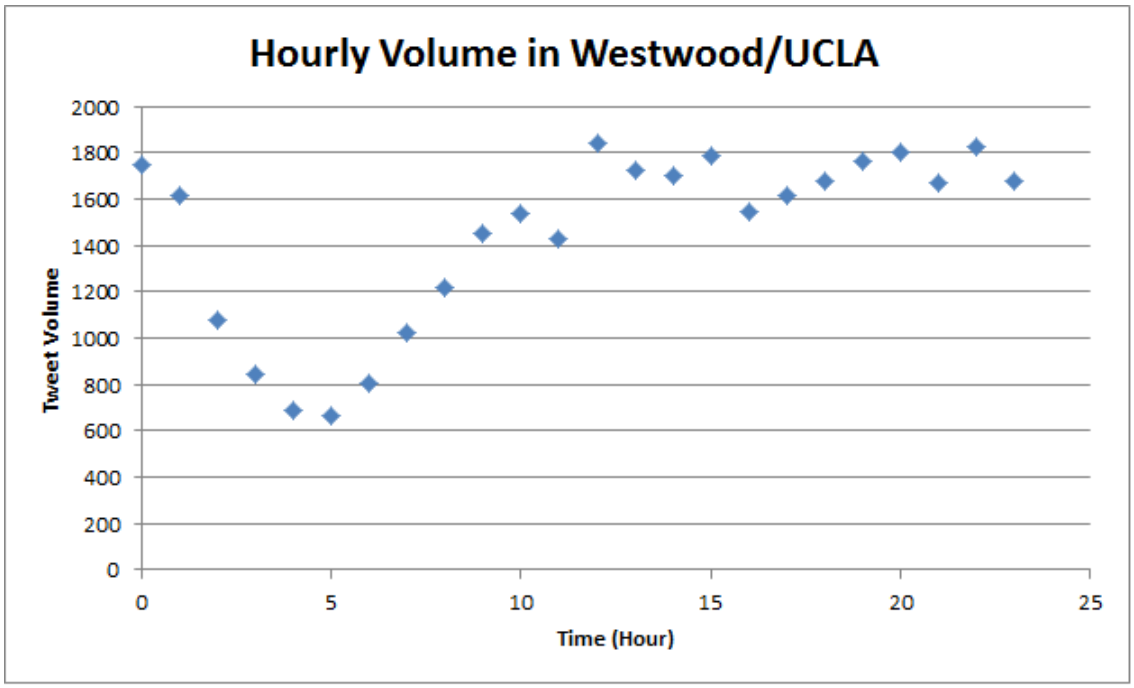

Figure 5.6: Hourly volume of tweets in Westwood/UCLA.

that occur in quick succession. It is then necessary to attach an uncertainty based on a users history of status updates. This uncertainty is a rough estimate of where a user will be based on their history. Using this information, we are able to answer the questions of where a user is traveling and how long they are staying in a given location.

Since it is impossible to obtain perfect certainty with the data, there must always be a degree of uncertainty attached to the results. This degree of uncertainty is dependent on how much data is available at the time of measurement. When looking at historic data for a user, it is possible to determine some patterns about the users movement. Since we have large quantities of historic user data, we start by analyzing data during certain hours of the day and certain days of the week. We find that when looking through this data, that a high degree of cluster-

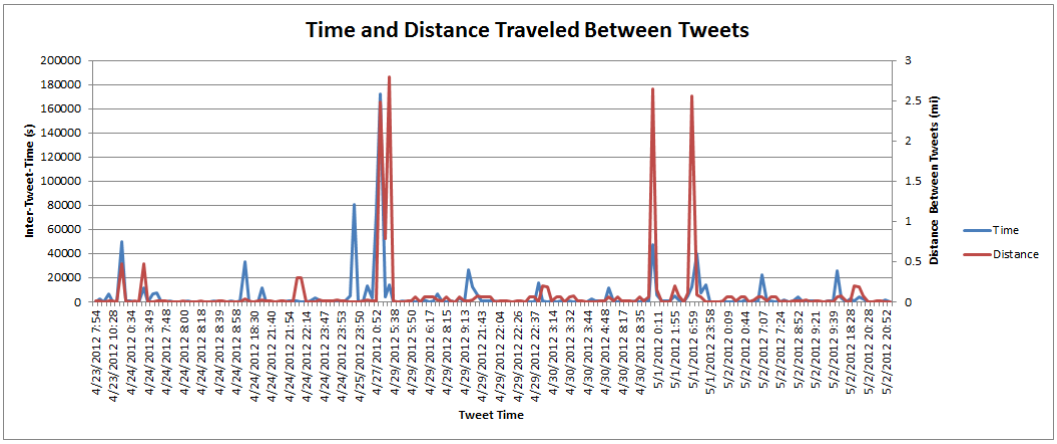

Figure 5.7: Graph of inter-tweet time and distance traveled between tweets.

Location	Tweets/Min	Average ITT	Median ITT	Average Dist
Macao	0.1768	428.6701	146	
City of Dreams	0.0876	686	177	
Heavy Users	0.0207	4530.896	308	0.0936
Medium Users	0.00126	47918.03	9961	19.1223
Light Users	0.000165	409082.6	63014	10.4356

Table 5.1: Table of inter-tweet times (ITT) and average distance between tweets.

ing occurs. This indicates that users tend to post status updates from the same locations. This reinforces the previous determination that users tend to have a single or few favorite locations to post from. We look at the map corresponding to a student at UCLA. We examine the users tweets on Mondays at several intervals during the day. In Figures 5.8, 5.9, and 5.10 we see strong clustering of tweets between weeks indicating that the users motion is quite regular. The points on the map indicate the number of times that user 1 has posted a status update at the location. These status updates are from several different Mondays at 8am, 10pm,

and 12pm. Similarly, we see in Figures 5.11, 5.12, and 5.13 that clustering occurs as well. These figures are from Saturdays at 11am, 12pm, and 1pm over a period of five weeks. Although it is difficult in user 2 to determine where they will go at 12pm, we can infer that they stay at their first location until then and only stay at these locations for one hour. The clustering of points over time is what allows us to reduce the uncertainty of the measurements. By analyzing points over a long period of time to determine patterns, we can probabilistically determine how long a user will stay in one location. Likewise, if the clustering is high, we can determine that the user will go to that location at the same time of day at the same time of week with a high degree of probability.

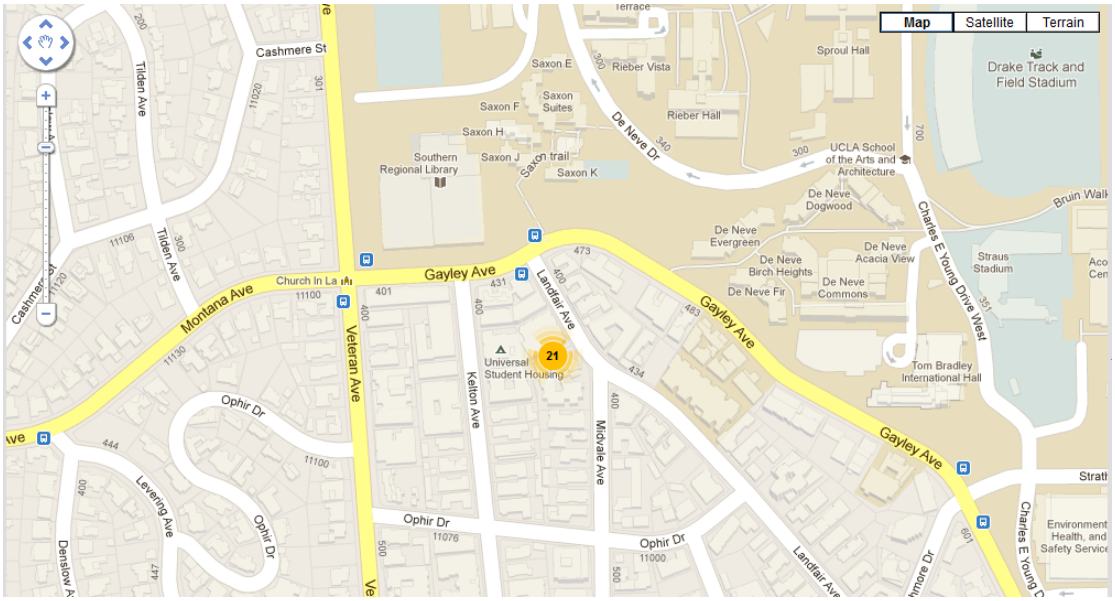

Figure 5.8: All tweets by user 1 at 8am on any given Monday.

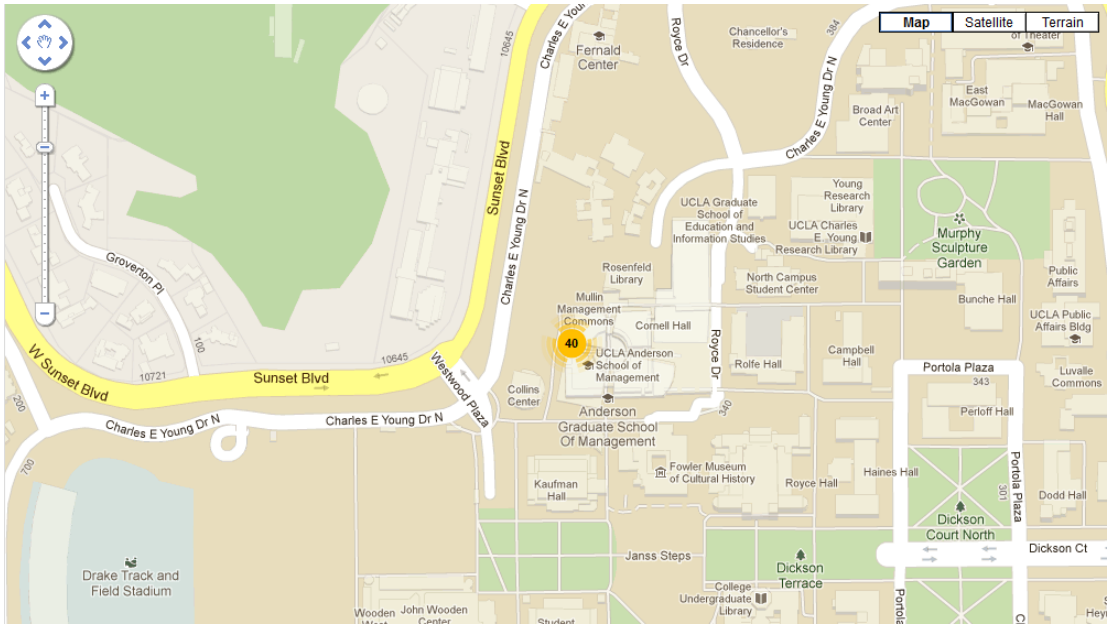

Figure 5.9: All tweets by user 1 at 10am on any given Monday.

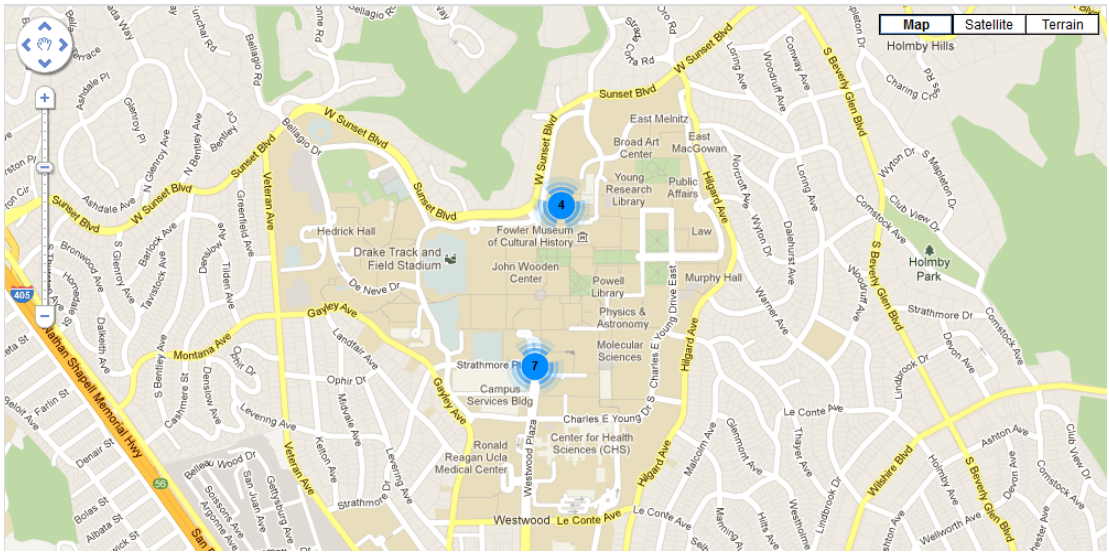

Figure 5.10: All tweets by user 1 at 12pm on any given Monday.

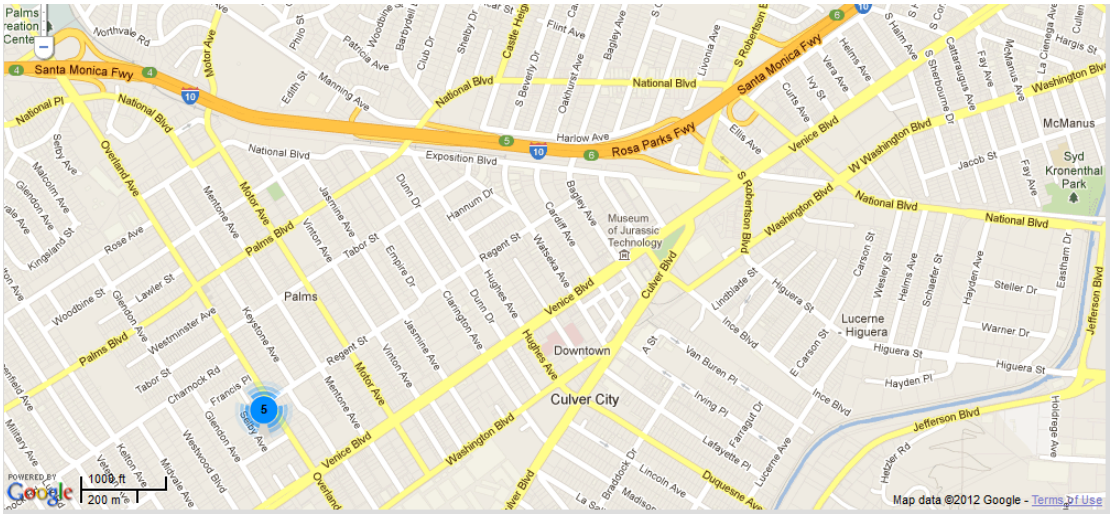

Figure 5.11: All tweets by user 2 at 11am on any given Saturday.

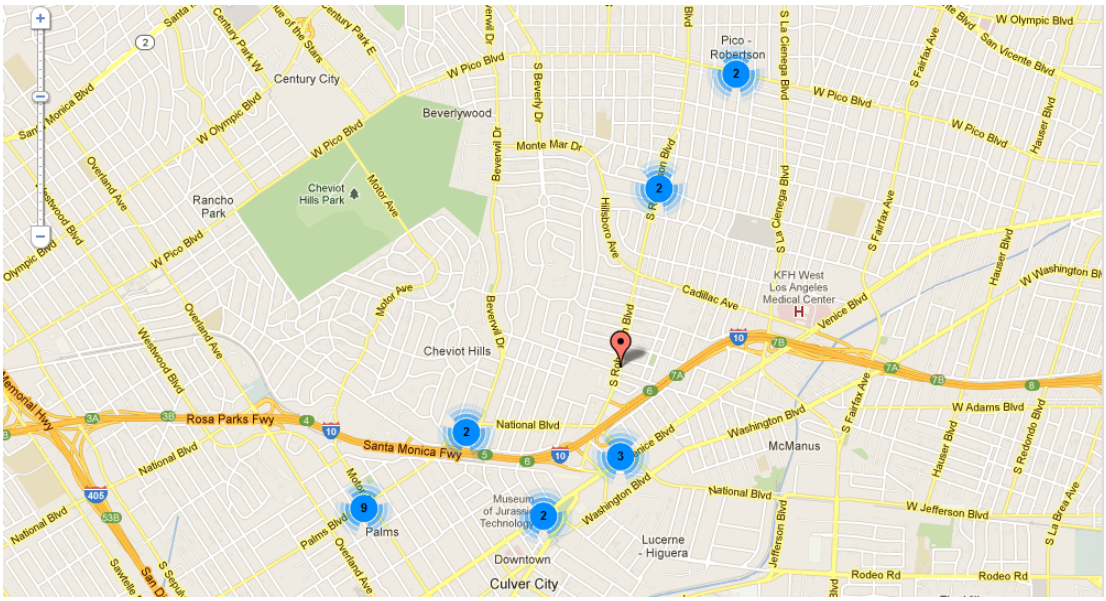

Figure 5.12: All tweets by user 2 at 12pm on any given Saturday.

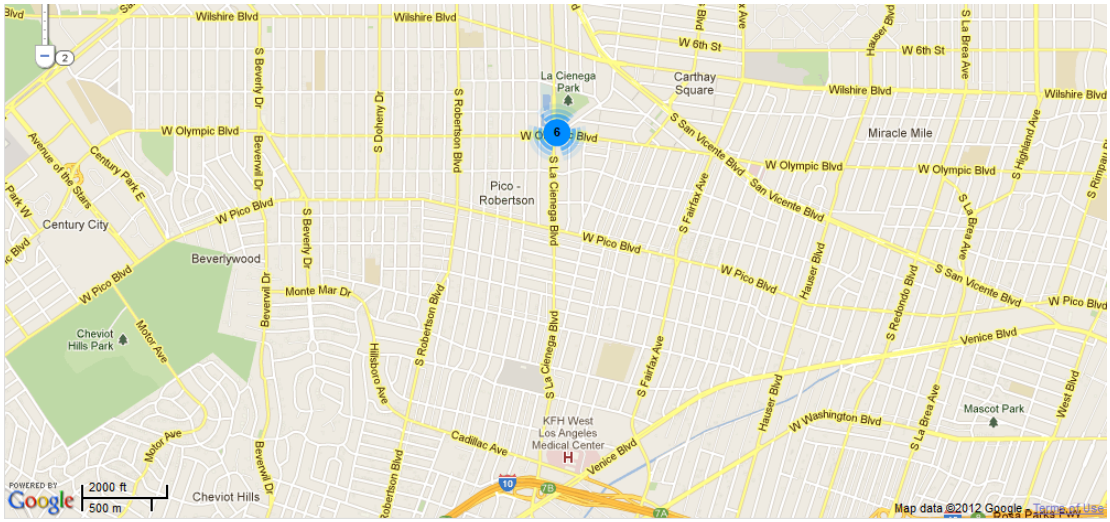

Figure 5.13: All tweets by user 2 at 1pm on any given Saturday.

CHAPTER 6

Implementation and Experiments

In our experiments, we apply our social network analysis model to real life applications. Analyzing human mobility in geo-location enabled online social networking websites has many important real world applications. We apply our model to look at pollution exposure using the Vehicular Sensor Network in the city of Macao in China. We also discuss possible applications in law enforcement, commercial advertising, and medical systems.

6.1 Applications

There are several applications that benefit from an analysis of human mobility in online social networks. These applications rely on knowing where people go and when they arrive. Commercial applications such as recommendation engines are particularly well suited for such knowledge of human mobility. A recommendation engine that relies on mobility information can better target advertisements and special discounts to people who tend to be in those locations frequently. These applications are very beneficial to small businesses who can target users who frequent popular venues nearby their business. Since we generate high probability locations for users, a recommendation engine can take these locations and target

users who will probably be there at specific times of day. This kind of recommendation engine is both more powerful and more accurate than those currently in place by many advertising firms.

Law enforcement applications can also benefit greatly from human mobility information. Law enforcement officials can use this information to pursue suspects in a criminal case, find fugitives, and curb violence. By analyzing the social trends of known offenders, law enforcement officials can leverage the power of our application to find out where they will go next and how long they tend to stay still. There have been reports of law enforcement officials using social networks to track suspects and our application provides an accurate and robust system for law enforcement officials.

Lastly, medical applications can greatly benefit from these applications. We address exposure to harmful pollutants in the air with the Vehicular Sensor Network. This data can help prevent asthma and other respiratory illnesses by alerting users to nearby pollution hot spots. Similar applications have also been used to track the spread of disease through online social network information. Such data is vital in helping to improve the quality of life in individuals who use social networking websites. As more users begin to use online social networks and take advantage of the geo-location features in these web sites, more users can become aware of potential health hazards in areas frequented in their daily routine.

6.2 Experiment and Results

With the help of the Macao Polytechnic Institute in the city of Macao in China, we use the Vehicular Sensor Network to monitor pollution in the city of Macao. The Vehicular Sensor Network consists of a fleet of vehicles with sensors attached to an on-board computer. These sensors monitor different aspects of the vehicle as well as the surrounding environment. Currently, we use the carbon dioxide sensors attached to the vehicles to produce a map of carbon dioxide concentrations throughout the city. This map is constantly updating throughout the day as the vehicles collect more data. When the vehicles return to the university, they upload the data to several systems which download and process the sensor information. Figure 6.1 shows sample data gathered from the vehicles.

Figure 6.1: CO2 data gathered from the Vehicular Sensor Network.

While the vehicles are collecting physical data from Macao, our application UCLASocial, is collecting data from several online social networking websites. This data is stored and replicated on several servers and database systems. The

data we collect for analysis purposes is from the same areas that the Vehicular Sensor Network is running. We collect data from here in order to determine what the people themselves are doing in these areas. Once we have gathered enough data, we can make inferences about where people go and how long they stay.

By leveraging the data gathered from the Vehicular Sensor Network and combining it with the data gathered by UCLASocial, we combine the data to combine the physical data with the data from cyberspace. We use this data to estimate the exposure to carbon dioxide of people living in these areas. With more sensors on the vehicles, the same method can be applied to volatile organic compounds and other harmful pollutants in the air. These results are important in educating people as to how much exposure they receive as well as improving the quality of life of individuals by alerting them to possible hot spots of pollution to avoid.

CHAPTER 7

Conclusion and Future Work

Social networking is more than just an online activity, it has become deeply ingrained into society. Millions of people worldwide post to online social networks regularly. As more people begin to post more and divulge more about their lives and daily habits, much can be learned about human social behavior by analyzing patterns and trends that occur in online social networks. In this thesis, we collected data from several online social networks, specifically those with geo-location features. Using this data, we were able to determine trends and patterns that formed in user behavior. These trends allow us to answer the questions of where people are going and how long they stay.

Geo-location enabled online social networks are a fairly new phenomenon. The users that use and post regularly to these web sites are divulging an unprecedented amount of private information about their daily lives. We can analyze the movement of these individuals to gain a lot of insight into human mobility and motion patterns in individuals. From analyzing those users who use online social networking websites frequently, we found several patterns that occur between users. We found high degrees of tweet location clustering over several months of data analysis. Users tend to post updates from a few familiar locations. Users regularly return to these locations in a predictable manner. By analyzing tweets over

several months by hour and day of the week, the clustering holds. This shows that user mobility in online social networks is fairly predictable and regular.

In analyzing how long a person stays in one location, we examined the time between status updates for our population. We found that users tend to post in large bursts and then take long breaks which may result in a change of location. By analyzing the time of these large bursts, we can calculate how long they were at the location. However, there is still uncertainty in this result. This is caused by travel time between locations and possible delays in user updates. By analyzing the status update patterns over a period of time, we are able to determine the probability that a user will stay at that location. This allows us to reduce the uncertainty of the measurement since we are able to attach all of the users historic data to the measurement as well.

When we have the answers to how long a user stays at a location and where they travel to, there are many applications that can be applied to this information. Pollution in major cities is a large problem and is getting worse in recent years. In the city of Macao in China, we used the Vehicular Sensor Network created by the Macao Polytechnic Institute to gather a live map of carbon dioxide concentrations around the city. We use this map in concert with the social network data to calculate exposure to pollutants in the city. Having the knowledge of your exposure to pollutants is very important. This knowledge can improve the quality of life for many people worldwide by alerting them to dangerous levels of pollution.

Online social networks, especially those with geo-location enabled features are not without drawbacks. Privacy is a major concern in online social networks and on the Internet in general. As users post more about their daily lives, their

private information begins to circulate around the web at incredibly high rates. Most social networking websites have privacy features that users can configure, however, the privacy features are either not enough or users simply do not enable them. There are measures that users can take to help preserve their privacy. Users can obfuscate their information, however, this can only go so far as to protecting their identity. The best thing users of online social networking sites can do is educate themselves to the dangers of posting private information online.

Online social networks are an excellent resource and supplement for our lives. They help us stay connected with friends, share interesting facts about our days, and allow us to connect with far more individuals than ever before. Social networks are a rich source of data for statistical analysis as well. There are countless applications to social network data ranging from medical to law enforcement to commercial uses. However, privacy is a huge concern with social networks as people tend to post so much about their private lives that they are no longer private anymore. Social networks have grown very large over the recent years and are here to stay. The APIs provided by developers of social networks provide a portal to an unprecedented amount of knowledge. Harnessing the vast knowledge of online social networks can prove to be the next gold rush of our generation.

References

[BW92] P. Bourdieu and L. Wacquant. *An Invitation to Reflexive Sociology. Chicago*. University of Chicago Press, 1992.

[CDF11] Salvatore A. Catanese, Pasquale De Meo, Emilio Ferrara, Giacomo Fiumara, and Alessandro Provetti. "Crawling Facebook for social network analysis purposes." In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, WIMS '11, pp. 52:1–52:8, New York, NY, USA, 2011. ACM.

[CML11] 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*, KDD '11, pp. 1082–1090, New York, NY, USA, 2011. ACM.

[com] "comScore." http://www.comscore.com/.

[CPA] "The Comprehensive Perl Archive Network." http://www.cpan.org/.

[ESL07] Nicole B. Ellison, Charles Steinfield, and Cliff Lampe. "The Benefits of Facebook Friends: Social Capital and College Students Use of Online Social Network Sites." *Journal of Computer-Mediated Communication*, **12**(4):1143–1168, 2007.

[GA05] Ralph Gross and Alessandro Acquisti. "Information revelation and privacy in online social networks." In *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, WPES '05, pp. 71–80, New York, NY, USA, 2005. ACM.

[Gil11] Felix Gillette. "The Rise and Inglorious Fall of Myspace." *Bloomberg Businessweek*, 2011.

[HH97] Michael Hauben and Ronda Hauben. *Netizens: on the history and impact of Usenet and the Internet*. IEEE Computer Society Press, Los Alamitos, CA, USA, 1997.

[HP09] A. Hughes and L. Palen. "Twitter adoption and use in mass convergence and emergency events." In *International Journal of Emergency Management*, volume 6, January 2009.

[Joi08] Adam N. Joinson. "Looking at, looking up or keeping up with people?: motives and use of facebook." In *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, CHI '08, pp. 1027–1036, New York, NY, USA, 2008. ACM.

[KKN07] Jussi Kuittinen, Annakaisa Kultima, Johannes Niemelä, and Janne Paavilainen. "Casual games discussion." In *Proceedings of the 2007 conference on Future Play*, Future Play '07, pp. 105–112, New York, NY, USA, 2007. ACM.

[LC10] V. Lampos and N. Cristianini. "Tracking the flu pandemic by monitoring the social web." In *Cognitive Information Processing (CIP), 2010 2nd International Workshop on*, pp. 411 –416, june 2010.

[LCC08] Ieng-Fat Lam, Kuan-Ta Chen, and Ling-Jyh Chen. "Involuntary Information Leakage in Social Network Services." In *Proceedings of the 3rd International Workshop on Security: Advances in Information and Computer Security*, IWSEC '08, pp. 167–183, Berlin, Heidelberg, 2008. Springer-Verlag.

[LMG09] Uichin Lee, E. Magistretti, M. Gerla, P. Bellavista, and A. Corradi. "Dissemination and Harvesting of Urban Data Using Vehicular Sensing Platforms." *Vehicular Technology, IEEE Transactions on*, **58**(2):882 –901, feb. 2009.

[LW02] Lee and Brent Ware. *Open Source Development with LAMP: Using Linux, Apache, MySQL and PHP*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2002.

[Mys] "MySQL." http://www.mysql.com/.

[Oau] "OAuth." http://oauth.net/.

[Per] "Perl Dancer." http://perldancer.org/.

[RFB11] C. Ruiz Vicente, D. Freni, C. Bettini, and C.S. Jensen. "Location-Related Privacy in Geo-Social Networks." *Internet Computing, IEEE*, **15**(3):20 –27, may-june 2011.

[S10] Venkatraman S. "Social Networking Technology as a Business Tool." In *Allied Academies International Conference Proceedings, Fall 2010*, 2010.

[Sce11] Salvatore Scellato. ""Beyond the social web: the geo-social revolution" by Salvatore Scellato with Ching-man Au Yeung as Coordinator." *SIGWEB Newsl.*, (Autumn):5:1–5:5, September 2011.

[Sha] Brian Shaw. "Understanding human mobility with machine learning and a billion check-ins." http://engineering.foursquare.com/2011/10/25/understanding-human-mobility-with-machine-learning-and-a-billion-check-ins/.

[SMM10] 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*, WOSN'10, pp. 8–8, Berkeley, CA, USA, 2010. USENIX Association.

[SOM10] 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*, WWW '10, pp. 851–860, New York, NY, USA, 2010. ACM.

[Ven] "Mappr." http://stamen.com/projects/mappr/.

[Wara] Pete Warden. "How I Got Sued by Facebook." http://petewarden.typepad.com/searchbrowser/2010/04/how-i-got-sued-by-facebook.html.

[Warb] Pete Warden. "How to Split Up the US." http://petewarden.typepad.com/searchbrowser/2010/02/how-to-split-up-the-us.html.

[YYL10] 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*, GIS '10, pp. 458–461, New York, NY, USA, 2010. ACM.