# LOCATION SHARING BEHAVIOR: ANALYSIS AND MODELS

A Thesis

by

HIMANSHU BARTHWAL

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

MASTER OF SCIENCE

| | |
|---|---|
| Chair of Committee, | James Caverlee |
| Co-Chair of Committee, | Richard Furuta |
| Committee Member, | Takashi Yamauchi |
| Head of Department, | Dilma Da Silva |

August 2015

Major Subject: Computer Engineering

# ABSTRACT

A range of mobile applications allow individuals to create geo-located content using location-based services like Foursquare, Facebook and Twitter. This location-based sharing creates new opportunities for users to associate their life events to geographic locations, e.g. *places they have been* feature in Facebook and share their locations in a social context with their connections on the social network to inform them about their whereabouts. But these opportunities present a clear risk to user privacy and location privacy. And yet, many users continue to voluntarily share their location information. In this thesis, we aim to study the factors impacting location sharing behavior toward providing a foundation for future adaptive location privacy systems which can help users decide whether it is safe to share their location or not. Concretely, we study a unique Twitter-based dataset of (i) users who always share their location, (ii) users who never share their location, and (iii) users who selectively share their location. We conduct a data-driven analysis of location sharing via multiple factors including the time of the Tweet, the content of Tweet, and user profile features. Based on this data-driven analysis, we investigate whether we can predict whether a Tweet will be tagged by a user with geo-location or not, a key step for enabling an adaptive location privacy system. We create a global classifier for all users to uncover the common driving factors for location sharing. We also build per-user individual classifiers to improve the prediction performance and to view the users in a spectrum of predictability. We achieve an accuracy of 70% for the global classifier and an accuracy greater than 90% for more than 60% of users. We observe that features like the users social status, the source of the Tweet, whether the Tweet has a mention or not, and the textual content of the Tweet are the most important features. These observations imply that users are conscious of their online visibility and social connections while geo-locating and also that the usage of mobile devices promotes location sharing. We also conclude that most users are highly predictable in terms of their location sharing behavior and thus our work creates a substantial groundwork for future data-driven location privacy systems.

# ACKNOWLEDGEMENTS

I am grateful to my advisor, Dr. James Caverlee, who gave me all the freedom to pursue my interest and directed me throughout the process of this research work.

I am thankful to my co-chair, Dr. Richard Furuta, and committee member, Dr. Takashi Yamauchi, for their valuable feedback on my work.

I take this opportunity to thank my lab members, Cheng Cao and Haokai Lu, whose active participation in discussions with me helped me to proceed in the right direction.

# TABLE OF CONTENTS

# LIST OF FIGURES

FIGURE Page

# LIST OF TABLES

# 1. INTRODUCTION

In the last five years, there has been a tremendous increase in the amount of geo-located content on the web and mobile services. Applications like Foursquare, Facebook and Twitter allow users to create content along with geolocating it. Figure 1.1 shows an illustration of the geo-location feature for Twitter.

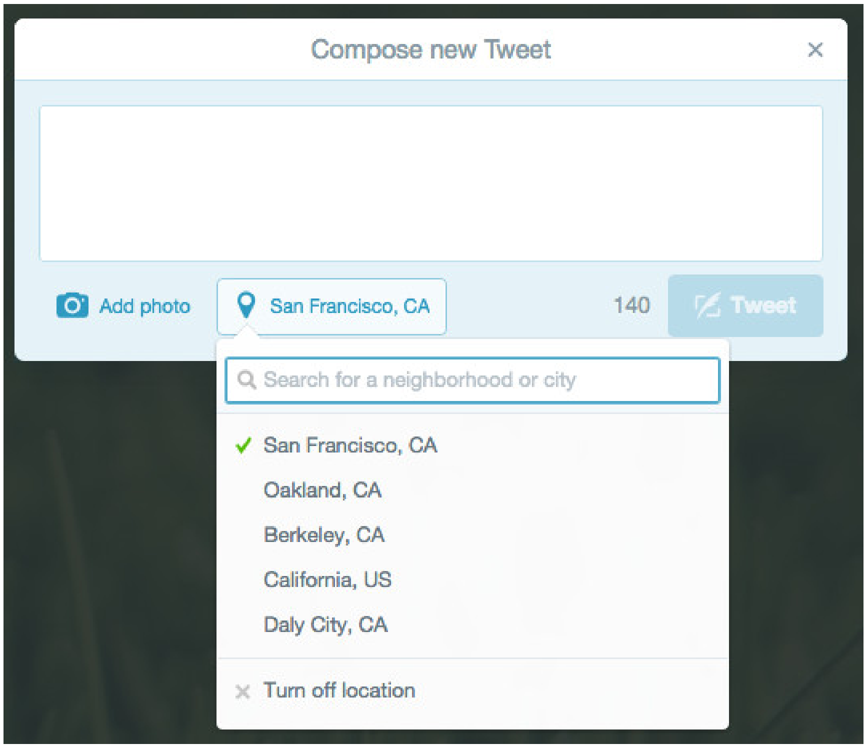

Figure 1.1: Geo-location feature in Twitter.

As shown in the figure, a user can post a Tweet while associating a geo-location with the Tweet, thus associating the user and the message with a particular location. In general, the level of granularity of a geo-location can be a latitude-longitude coordinate, a neighborhood or venue, or a city. The number of users of these location sharing applications is growing very rapidly. For instance, Facebook has 1.2 billion users and Twitter has more than 200 million users. Arguably, these applications are responsible for most of the geo-located content generated by individual users as well as organizations and therefore provides a great opportunity to analyse and understand the location sharing behaviour of users, which can potentially be useful for:

- Building new models of how and why users decide to share their location. Under what circumstances do users share their location? And when do they choose not to? And do these decisions vary based on the characteristics of individual users?
- Developing location privacy protection mechanism based on a data-driven investigation of location sharing in the context of social media. For example, can we advise users when they might elect to share their location or warn users when location sharing is activated (but perhaps not preferred)?
- Developing an application programming interface which can be used by location-based services to provide push notifications to users. For example, if a user is fond of sharing location while having lunch then it is may be likely that she wants to be notified about restaurants in her vicinity during lunch time.

In this thesis, we aim to investigate users location sharing behavior in public social media – in particular via a study of Twitter, rather than Facebook or Foursquare, primarily because of the following two reasons:

- The web service based Twitter sampling API provides unbiased Tweet data for free.

- The nature of the data is public and therefore the licensing terms will not interfere with our research work.

The Twitter community represents a variety of users representing different demographics and generates a lot of content which is either geo-located or not. We begin a study here into these users' motivation of geo-locating their Tweets. Specifically, we investigate whether the content (Tweets) that the users geo-locate differs from the content which they do not geo-locate. And if it does differ, then what are the prime attributes of the content which represent the contrast between these two (geo-located and non geo-located) categories.

We analyze the content of the Tweets generated by the users by using LIWC [31] labels based tagging on the Tweets text. We also look at the attributes of the users, i.e. their social status (ratio of number of followers and number of following users), the devices being used to generate the content, the temporal features, i.e. the time of generating the content (Tweet), mentions in the Tweet, and so forth.

In contrast, most recent research relies on surveys over a small population rather than the large-scale real world data generated by users via location sharing services. The closest data-driven analysis of location sharing was done by Bigwood et al. [8] however, their focus was primarily on the demographics of the users of a particular system (now obsolete) and did not attempt to analyze the content generated by the users in the social media with a perspective of understanding their location sharing behaviour. The most notable contributions of our work can be summarized as follows:

- We give insight into the usage of location sharing feature in social media, based on our analysis of a real-world Twitter dataset. Specifically, we answer questions like:
    - What *kind of content* is associated with geo-location in social media?

  - How does geo located content differ from non geo-located content?
  - How do the devices (mobile vs desktop) influence location sharing behavior?
  - How does the social status of a user relate to her location sharing behavior?

- We model the location sharing behavior of Twitter users using content-based and profile-based features using a decision tree classifier and rank the features according to their importances.

The next section presents the discussion of related prior research work. In section 3 we discuss the data set and our analysis approach. Section 4 describes our approach for building the decision tree based classifier and the results obtained. Finally, Section 5 summarizes our conclusions and presents directions for future work.

# 2. RELATED WORK

The geo-location feature has been used increasingly in online social networks and social media applications. Cramer et al. [11], Page et al. [13], Lindqvist et al [12] and Consolvo et al. [20] provide in depth, subjective studies of user perception about location sharing based on detailed interviews with users of location sharing applications. Location sharing has mostly been studied in the the context of location privacy of the users, for example, Barkhaus et al. [6] conducted a study of users' privacy concerns with respect to location based services. Research work by Minch et al. [18] also gives insight into the various aspects of privacy issues related to mobile device based location sharing. Bigwood et al. [8] have presented an initial investigation into the predictability of users' location-sharing privacy preferences in mobile social networks. Their work was based on a survey conducted with the help of 80 participants by asking them questions regarding their location sharing behavior in Facebook. Nan et al. [31] found that the characteristics of users' privacy protection behavior is correlated with their age, gender, mobility, and geographic region.

Many aspects aspects of location sharing have been studied throughout the literature. For example, Tsai et al. [7] studied that how feedback in location sharing impacts the behavior of the users. Lin et al. [2] suggest that both users' expectation and the purpose of why sensitive resources are used have a major impact on users' subjective feelings and their trust decisions. Another major finding is that properly informing users of the purpose of resource access can ease users' location privacy concerns to an extent. Hence, it becomes very important that the benefits of location privacy are provided to the users while alleviating the location privacy risk.

Toch et al. [4] suggest that location sharing privacy settings that enable users to restrict location disclosure to particular times and places, seem to play an important role in capturing peoples privacy preferences, especially those of more mobile users. Similarly, Sadeh et al.[23] have also contributed toward capturing location privacy preferences of users of location sharing services. Benisch et al. [5] found that more complex privacy-setting types, such as those that allow users to specify both locations and times at which they are willing to share, were significantly more accurate under a wide variety of assumptions. They also found that more complex setting types also generally lead to more sharing due to the fact that users generally tend to err on the safe side, and restrict access with simpler settings.

Arguably, it can be said that improving the usability of location privacy systems is one of the key factors which needs to be addressed in order to improve location-based services both from a business and a user perspective. There has been extensive research in the field of location privacy which deals with different techniques of preserving location privacy based on approach like k-anonymity, p-sensitivity and l-diversity etc. Wernke et al. [3] and Krumm et al. [17] provide an overview of different kinds of privacy attacks and the methods to tackle those attacks. While these techniques do help users to maintain their location privacy, these approaches often result in the degradation of the overall quality of location data and therefore many services which depend on location data suffer from the loss of quality. This is mainly because most of these approaches treat location data as geometric data and not as context rich data which is generally the case with the current location-based services in social media/network. There have also been research efforts which focus on the contextual aspect of location sharing data. For example, Zhao et al. used a collaborative filtering technique to predict users' location sharing preferences using a data set acquired from 40 users. Xie et al. [16] developed a location privacy preference recommendation algorithm based on the data collected from Amazon Mechanical Turk, recruiting around 1000 users for data collection. Oh et al. [14] use a decision tree approach to take inputs from the user about their location privacy preferences and then proposed a b-diversity algorithm to to protect the privacy of the user according to her contextual preference specified with the help of the decision tree. However, the sample size (n = 10) is very small and the semantics of the location are not learned from the data but are rather taken as an input from the user. Lee et al. [21] propose an approach for location privacy protecting techniques based on the location semantics from the LBS applications by performing cloaking on semantically heterogeneous locations. They combine the location semantics with the previously mentioned location privacy techniques like l-diversity and k-anonymity to cloak the location of the user.

# 3. ANALYSIS

In this chapter, we initiate our data-driven investigation of location sharing through an analysis of a large Twitter dataset. The overall objective of this section is to build a general understanding of location sharing behavior and the various factors influencing it.

We begin with an examination of the overall shape of data. For example: How many users are there in our dataset? How many geo-located and non-geo-located Tweets do we have? What are the different types of devices used for posting Tweets by the users? And so forth. Next, we categorize users on the basis of their location sharing behavior and also try to understand these categories from the perspective of the content (Tweets) generated by them and their temporal activity patterns. Then, we move our focus to *Selective Sharers* category and try to understand their location sharing behavior by asking specific questions and answering them on the basis of our data analysis.

## 3.1 Data Overview

We now provide an overview of the data that we use for our analysis. In all the plots below, the data used is 6 months Tweet data (from Jan 2013 to June 2013). It contains about 780 million Tweets generated by around 75 million Twitter users. These Tweets were downloaded by a continuously running crawler at Infolab [33] which makes web-service based calls to the Twitter Sampling API [32] hosted on Twitter cloud.

### *3.1.1 Geotagged vs. Non-Geotagged Tweets*

Our dataset contains Tweets of both geo-located and non geo-located category. As per the pie chart shown in the figure 3.1, the non geo-located Tweets constitute 98.5% of the Tweets which indicates that geo-location is a sparingly used feature in public social media like Twitter.

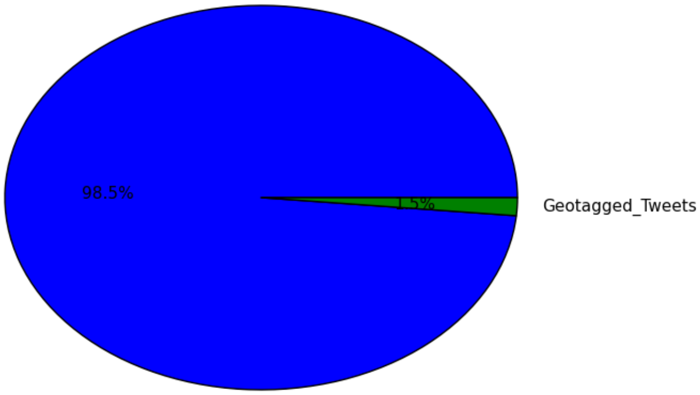

Figure 3.1: Geo-located vs. non geo-located Tweets in the dataset.

### *3.1.2 Tweets Distribution*

The following plot shows the log-log plot of the number of users and the number of Tweets. As we can see in figure 3.2, most users generate very few Tweets while a small number of users are responsible for generating a high number of Tweets.

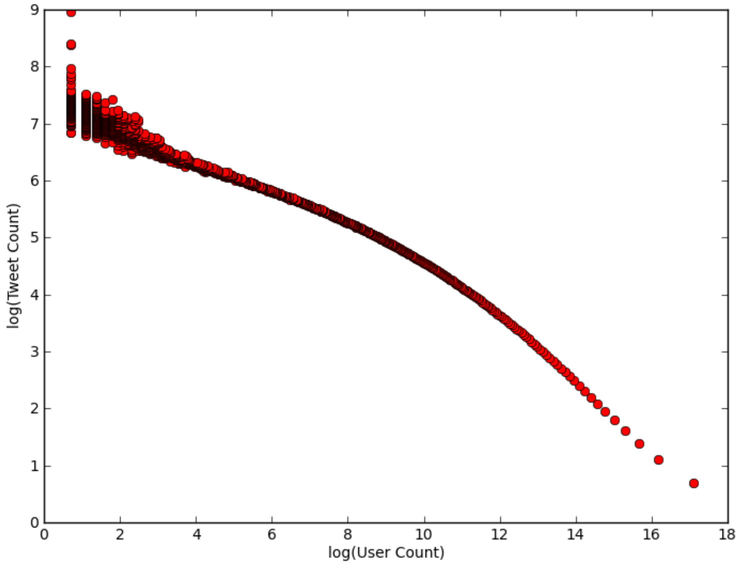

Figure 3.2: Log-log plot of users and Tweet count.

### *3.1.3 Tweet Source*

The plot in Figure 3.3 shows the different sources from which the Tweets are generated, implying that, at least 70 % of Tweets are generated from mobile devices.

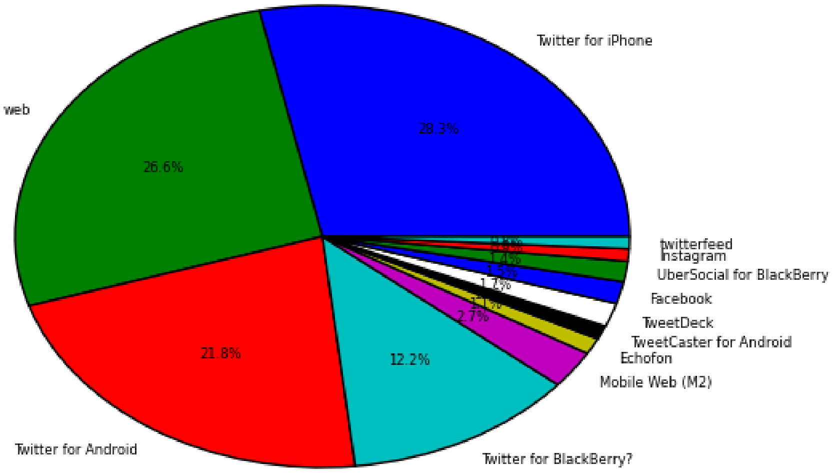

Figure 3.3: Figure showing the percentages of Tweets generated by various sources.

### *3.1.4 Places Distribution*

The Twitter geo-location feature provides the option to associate a certain location with a place type for example country, city, admin, neighborhood and *poi* (point of interest). The plot in figure 3.4 shows the percentage of the geo-located Tweets with different place types.

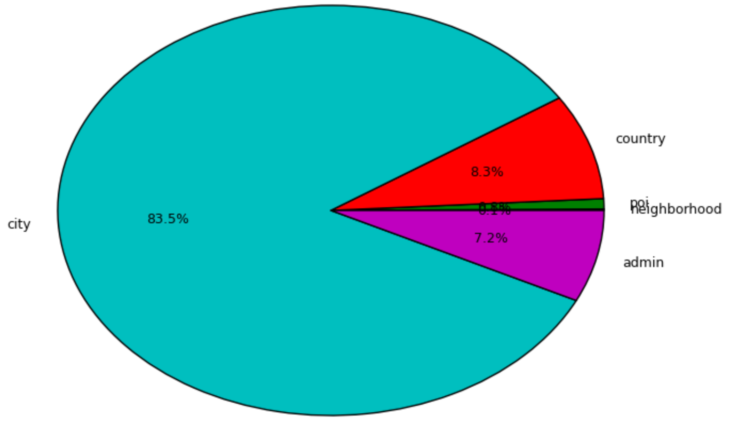

Figure 3.4: Figure showing the percentages of place types in the Tweets.

### *3.1.5 Geo-located Tweets Distribution*

In the Figure 3.5, the x axis represents the percentage of users to the total user base who share geotagged content, and the y axis represents the percentage of content of the user is geotagged. About 90% of users don't geo-locate their Tweets, while about less than 1% of users geo-locate all Tweets.

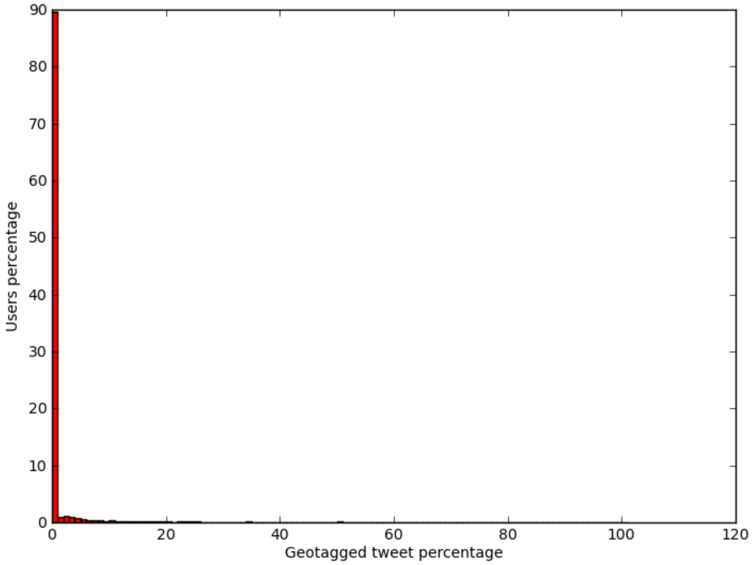

Figure 3.5: Geo-located Tweets percentage vs. users percentage

### *3.1.6 Observations*

- Most of the users (90%) are not comfortable in sharing any location data while only 1% share all of the locations in their Tweets. This observation is supported by:
    - research work by Kelley et al. [1] which presented an empirical study

  showing that if users are given only an optin/optout mechanism, a large percentage of users are unable to specify their true privacy preferences, and they stop using geo-location entirely.

  - the study conducted by Wagner et al. [10] in which participants indicated hesitation toward broadcasting their location and preferred sharing it on a need-to-know basis. Since Twitter follows a broadcast model, it is unsurprising to observe such user behavior.

- The people who do share locations are also very selective and therefore most of them end up sharing very few locations.

- Most of the Tweets (70%) are generated by mobile devices and rest (30%) are generated by other sources like web interface for non mobile devices.

- Most of the times – when sharing location information – the users are comfortable in sharing city level locations and they rarely tend to share neighborhood level locations which is in agreement with Wilson et al. [9] suggesting that most users think that sharing city level location is like sharing nothing at all.

## 3.2 Analyzing Location Sharing

In this section we analyse the various features of the Tweet and the users and their correlation with the location sharing behaviour of the users.

### *3.2.1 Categorizing Location Sharing Behavior*

Based on location sharing preferences, we can categorize users into three broad categories:

1. The users who share locations in all of the Tweets. We call them All Sharers.

2. The users who selectively share their locations in some of their Tweets. We call them Selective Sharers.

3. The users who do not share any of their locations. We call them Non Sharers.

In our dataset we have around 75 million *Non sharers*, 2.2 million *Selective sharers* and 450,000 *All Sharers*. Now we perform the following analysis:

1. We aggregate the Tweets of each user and considered this aggregated Tweet text as a single document $d$, which corresponds to a single user who can belong to any of the three categories mentioned above.

2. We calculate the mutual information of the words in these document with respect to the above three categories and then picked the top 200 informative words from each category similar to [24]. Mutual Information for word $w$ for user category $c$ is defined as:
   $MI\left(w,c\right) \;=\; p\left(w|c\right)p\left(c\right)log\left(\frac{p(w|c)}{p(w)}\right)$

Table 3.1 shows the most informative words for the three categories of the users. We can see that the All Sharers category is full of the names of places, cities, states and words representing weather conditions. We expect that in this category most of the accounts are driven by organizations like news channels, food businesses and the individual users who share all the locations and do not care much about their location privacy when they are Tweeting. The Non Sharers category is characterized by the account handle names which were suspended by Twitter due to violation of Twitter Rules [25]; Twitter stop words like RT, Follow etc; generated content via various applications like Facebook, Youtube etc; and famous Twitter accounts owned by individuals and organizations. However the Selective Sharers' group most closely resembles the language model of regular individual users because the most

Table 3.1: Table containing the most informative words for the All Sharers, Non Sharers and Selective Sharers.

| Users Category | Most Informative Words |
|---|---|
| All Sharers | Airport, Hotel, Cafe, Bar, Casino, Restaurant. Names of cities , states and other public places like Park, Gym, Street, Garden, Thunderstorm, Humidity, Weather, Storm, Freeze, Tornado, Cloudy |
| Selective Sharers | school, friends, love, hate, people, feel, girl, miss, sleep, morning, person, birthday, with, you, he, she, just, like, Iam, at, my |
| Non Sharers | RT, Follow, Suspended accounts like OMGFunniest, TheRrealTed, IbraheemMopelol. Celebrity accounts like FuelOnline, NialOfficial, Ipad, AndroidGames, Youtube, Google, Facebook |

informative words are almost similar to spoken English. We assume that *most* of these users accounts are not some organization accounts or some marketing campaign driven accounts. Therefore, we conclude that this category of users is dominated by the individual users sharing their daily life experiences.

### *3.2.2 Temporal Activity*

We also compare the three categories on the basis of their temporal activity patterns. In the figures 3.6, 3.7 and 3.8, each point on a particular plot represents the cumulative sum of the number of active users of that particular category on the given *day of week*. For the plots in figures 3.11, 3.9 and 3.10, we divided each day into eight parts such that each interval is three hours long and therefore each point represents the cumulative sum of the number of active users in a particular three hours window. The plots in the Figures 3.6, 3.7 and 3.8 shows the activity level on each day of the week for Non Sharers, All Sharers and Selective Sharers respectively.

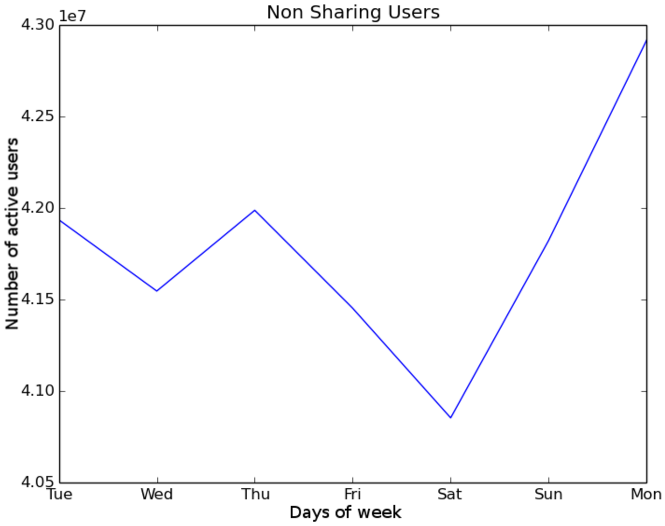

Figure 3.6: Activity level plot for Non Sharers.

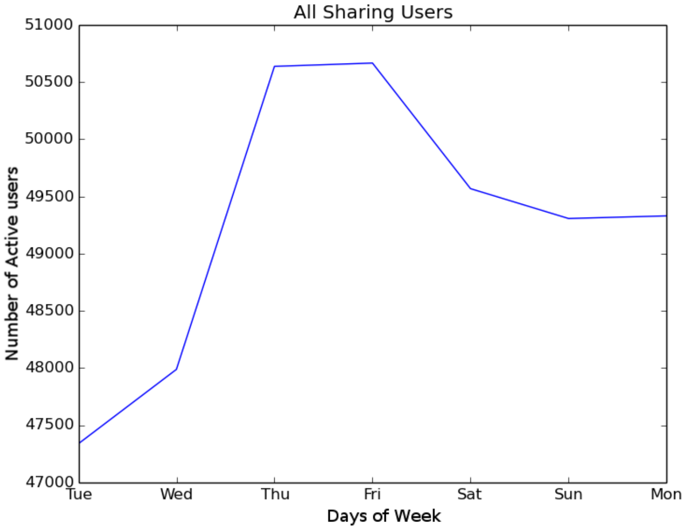

Figure 3.7: Activity level plot for All Sharers.

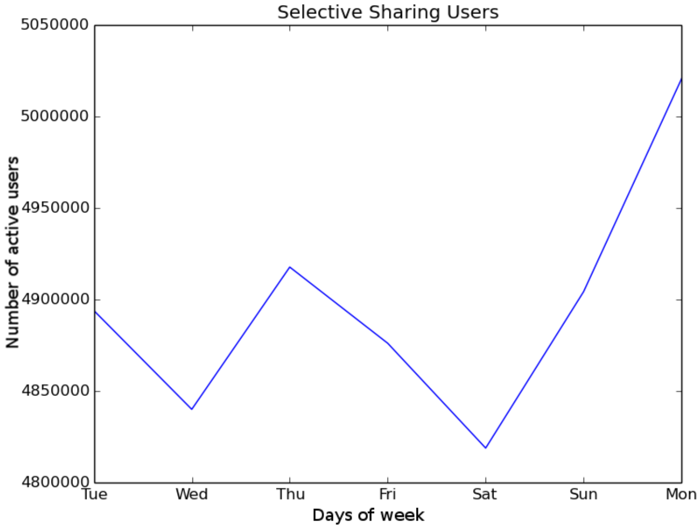

Figure 3.8: Activity level plot for Selective Sharers.

The plots in Figures 3.11, 3.9 and 3.10 show the activity level on each day of the week for each three hour window for Non Sharers, All Sharers and Selective Sharers respectively.

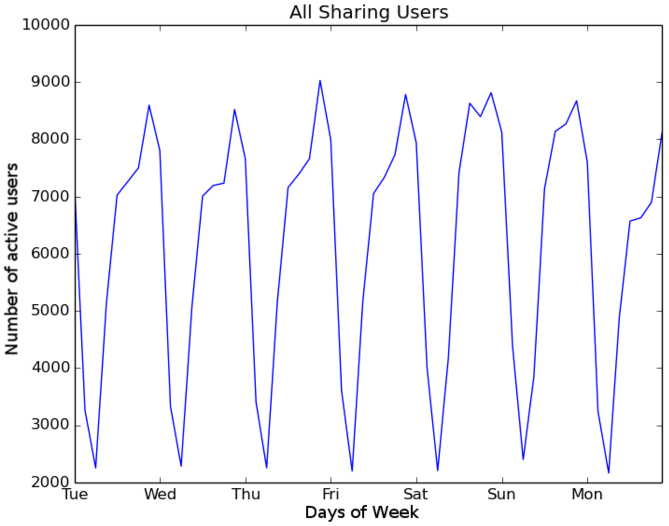

Figure 3.9: Activity level plot for All Sharers with three hour window.

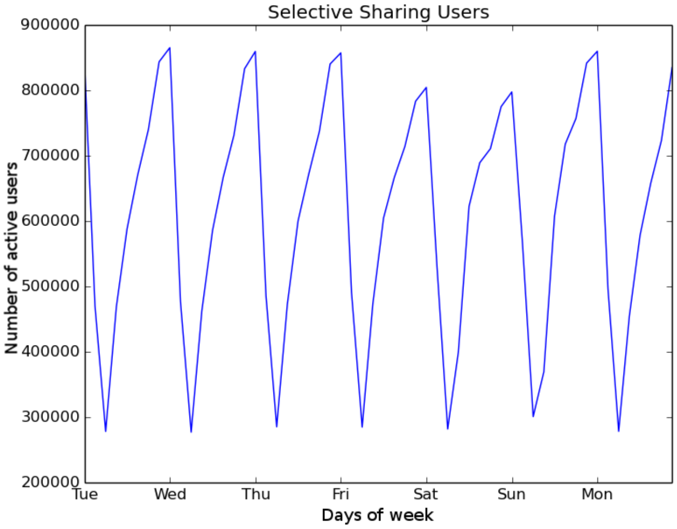

Figure 3.10: Activity level plot for Selective Sharers with three hour window.

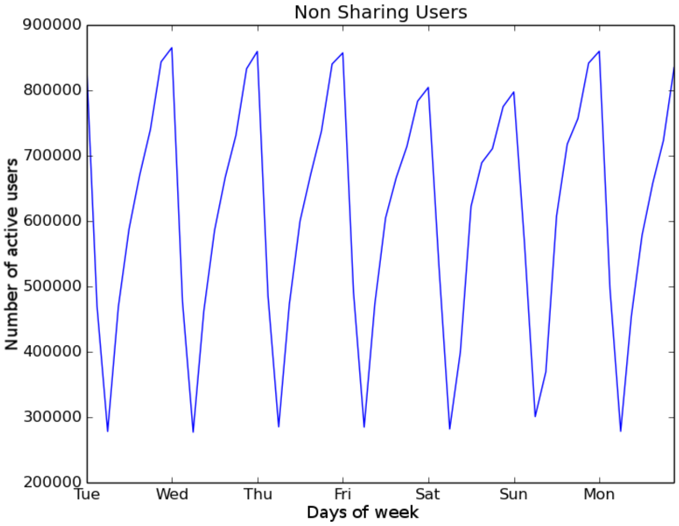

Figure 3.11: Activity level plot for Non Sharers with three hour window.

### *3.2.3 Observations*

1. We can see that the All Sharers particularly become active on weekends unlike the other two categories (Figures 3.6, 3.7 and 3.8)

2. Also we see a significant contrast in the activity level for All Sharers between daytime and evening which is not the same for the other two categories (Figures 3.11, 3.9 and 3.10)

3. The activity pattern of All Sharers changes on weekends significantly, and instead of one, two prominent activity peaks are observed.

### *3.2.4 Conclusion*

Since Non Sharers and Selective Sharers represent most of the Twitter users (99%) in our dataset, we argue that All Sharers are possibly involved in some focused advertisement or campaigning activity which is trying to target normal users to promote their products and services using geo-location during the weekends and the same applies for the daytime and evening activity patterns. Investigating this however is out of the scope of our work and we leave it as an open premise for future work.

For further analysis, we focus on Selective Sharers, primarily because:

1. The users in the Selective Sharers category post both geo-located and non geo-located Tweets which can help us understand the selective nature of the users in the context of location sharing.

2. Selective Sharers tend to have more keywords which seem to be generated by individual users and therefore we expect a lesser amount of automatically generated content.

For example, the commercial/organizational accounts like news channels tend to generate a lot of geo-located data in the All Sharers category, which we think is not suitable for the analysis that we intend to do with the data.The following section deals with studying the various aspects of the location sharing behaviour of the Selective Sharers.

## 3.3 Studying Selective Sharers

As described in the previous section, Selective Sharers are the Twitter users who share locations in some of their Tweets i.e. at least 1% and at maximum 99% of their Tweets are geo-located. These are the users whose location sharing behaviour can give us insight into the location sharing behaviour of regular individuals. In the next subsection we provide an overview of the data corresponding to Selective Sharers.

### *3.3.1 Data Overview*

For our study we only focus on English-speaking users: we use the language field in the users profile data to filter all the English-speaking users. In our dataset we have around 17.5 million Tweets in English, generated by 2.2 million Selective Sharers (again, who are users who share location in 1% to 99% of their Tweets). Around 82% of the Tweets are non geo-located and 18% are geo-located. In figure 3.12 the x axis represents the percentage of the users and the y axis represents the percentage of geo-located Tweets. Figure 3.13 shows the Tweet sources of the Selective Sharers.

### *3.3.2 Observations*

We can see that the users are reluctant to share a larger proportion of geo-located Tweets and most of them choose to share location with a very small percentage of Tweets. Also, the Selective Sharers prefer to use mobile as we can see that web source only accounts for 18% of all Tweets. In the following sections we ask specific research

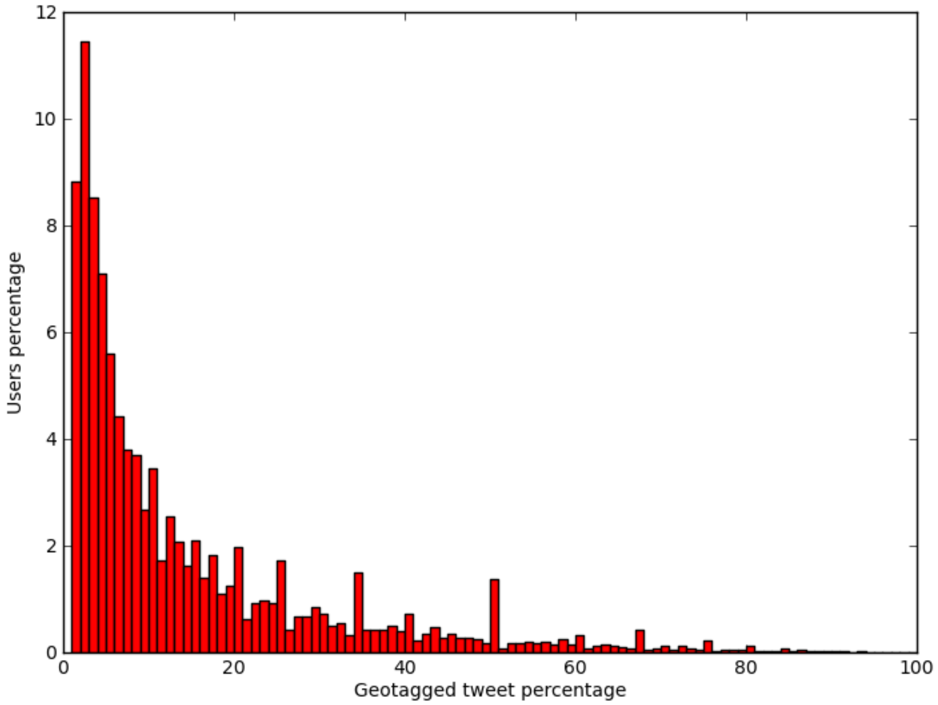

Figure 3.12: Geo-location percentages for Selective Sharers.

questions and then we discuss our analysis approach and the obtained results.

### *3.3.3 Is There Any Specific Context Associated with the Geo-located Tweets ?*

To understand the context of the Tweets we use word clusters [26] generated by CMU-ARK [27] lab. All the words in a Tweet are tagged by a cluster label. Then we order these cluster labels by the mutual information of these tags for geo-located Tweets. Mutual Information for word cluster label $l$ for Tweet category (i.e. geo-located or non geo-located) $c$ is defined as:

$MI\,(l, c) \;=\; p\,(l|c)\,p\,(c)\,log\left(\frac{p(l|c)}{p(l)}\right)$

The table 3.2 shows the various word clusters and their overall context. These are the top 50 word clusters obtained by arranging them in the order of their mutual

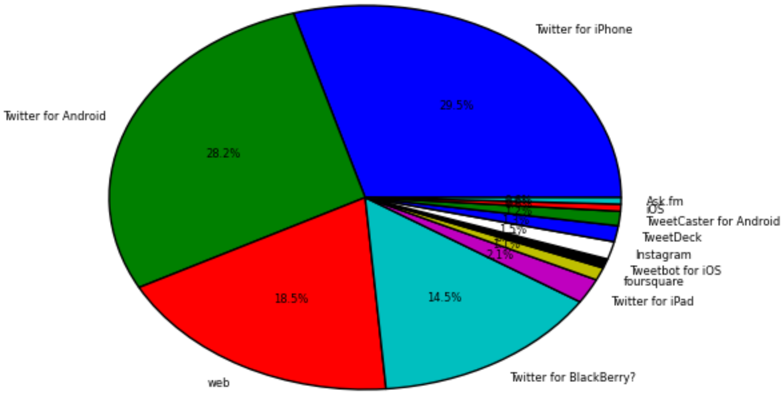

Figure 3.13: Tweet sources for Selective Sharers

information score for the geo-located category of Tweets. Each word cluster has anywhere between 1 to 5000 words and the table only depicts the top most frequent 5-10 words. Also, to avoid false positives we eliminate all the word clusters which do not have atleast 70 % of the words found in our dataset. Overall, the CMU-ARK word clusters represent 87 % of the vocabulary of our dataset.

We do the same analysis for the LIWC [28] categories and the top 10 LIWC labels are obtained as follows:

*space, i/them/her/she, ingest, preps, home, leisure, assent, past, work, relativ*

The meaning of all the LIWC categories is discussed in the table given on the LIWC website [29].

### *3.3.4 Conclusion*

We can see that the most informative word clusters and LIWC labels indicate that geo-located Tweets are often associated with individual users talking about various aspects of their daily lives like travel, food, activities and there is an associated emotional aspect to the geo-located Tweets. Our findings are in sync with the findings of research work by Tang et al. [19]. The word clusters shown in table 3.2 match with the Taxonomy discussed in their research work for place labels that includes both semantic and geographic place names.

### *3.3.5 Is There Any Correlation Between the Social Status of the Users and Their Likelihood of Sharing Location?*

As a crude approximation, the social status of the user is defined as the ratio of number followers of the user to the number of users that the given user is following. Formally, social status, $S = \frac{Number\, of\, followers}{Number\, of\, following}$ [24].

The location sharing likelihood is defined as the ratio of the number of geo-located Tweets to the total number of Tweets generated by the given user. The Location sharing likelihood is: $p = \frac{Number\, of\, geo-located Tweets}{Total\, number\, of\, Tweets}$.

The scatter plot in Figure 3.14 shows the users, represented as circles and the axes representing the social status (between 1 and 10) and the location sharing likelihood (between 0.01 and 0.99). The social status of the users represents their importance or status in the social network.

We have truncated the users with social status greater than 10 because:

- Firstly they are very few (less than .001 %) in comparison to the number of users with social status less than 10.
- Secondly, these users are outliers with respect to the overall distribution of the data and by eliminating them we can clearly see the ”hyperbolic” curve which shows how the users the distributed in the axes.

The users' social status and their likelihood of sharing location is weakly negatively correlated (Spearman correlation, $\rho = -0.3$). We can see that users with

Table 3.2: Table containing the most informative words in the geo-located Tweets for the Selective Sharers.

| | |
|---|---|
| Personal | at<br>i'am<br>@, @the, aht |
| Location / Venues | la, ny, washington, dc, downtown, seattle, ohio<br>park, market, center, court, hill, cafe, lounge<br>london, america, chicago, nyc, vegas, texas, india<br>etc, ca, co, pa, ga, inc, tx, va, fl, wa<br>beach, gym, airport, studio, mall, pool, hospital, hills<br>college, church, jersey, starbucks, midnight, target, youth<br>city, area, county, flu, island, league, region, fighter<br>country, hotel, storm, theme, garden, restaurant,<br>south, west, north, east, central, los, las, southern<br>street, st, bay, lake, river, bell, square, mountain<br>home, homee, hme, homeee, homeeee, home-<br>house, apartment, crib, closet, neighborhood, apt<br>state, king, bank, land, queen, university, sea<br>road, bus, train, field, plane, clock, boat, cab |
| Emotional | smh, jk, # fail, # random, # fact, smfh, # smh<br>!!, !!!, !!!!, !!!!!, !!!!!!, .!, !!!!!!!, !!!!!!!!, !!!!!!!!!, ..!<br>):, ]:, URL-minu.ws, )):, URL-ooyyo.com, ).:<br>:( , :/, -.- , :-\ , :-(, ":(", d:, :\| , :s<br>haha, hahaha, hehe, hahahaha, hahah, aha |
| Application generated | URL-myloc.me, URL-yfrog.com, URL-Tweetphoto.com<br>URL-, http, ht, htt, dvdrip, URL-t.c, URL-bit.l |
| Food | coffee, tea, beer, wine, juice, coke, beans, vodka<br>bar, box, ball, tree, table, corner, crowd<br>chocolate, chicken, cake, pizza, cheese, fish, milk<br>dinner, lunch, breakfast, brunch, dessert, supper |
| Temporal | night, nite, ”nights”, nigh, nites, nightt<br>day, dayy, nighter, workday, day-, dayyy, seriousness |
| Activity | shopping, swimming, ham, bowling, fishing<br>work, wrk, grub, werk, workk, workkk, work- |
| Others | pic, picture, vid, gif, screenshot, freerepublic<br>with, alongside, wtih, wih, w/all, wiht, woth, withe<br>was, ws, wuz, wass, waz, wus, wz, wasss<br>going, goign, goiing, reverting, goig, iwent, goint, gonig<br>good, gud, gd, goood, gooood, goooood, qood<br>american, national, chinese, international, global<br>#jobs, #tcot, #news, #job, #Tweetmyjobs, #music |

higher social status seem to be more reluctant to share their location. The users with very high social status are outliers and this is most probably because these types of accounts are rather for promotion or publicity of content and therefore the motivation of sharing location is different from that of an individual user.

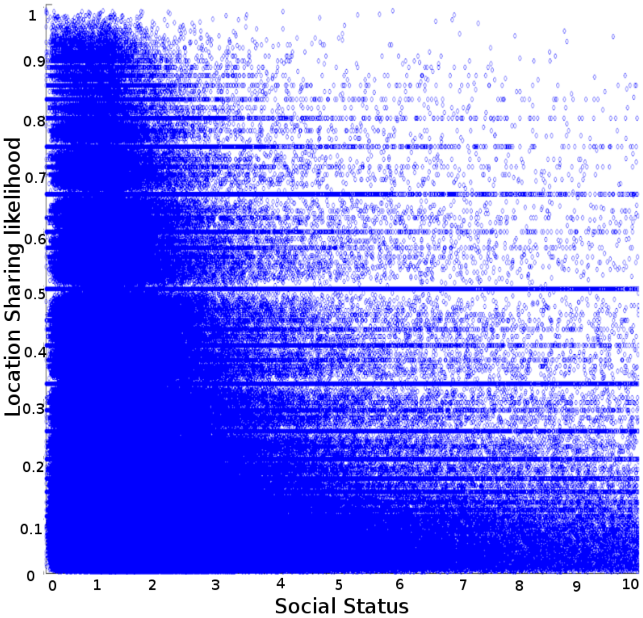

Figure 3.14: Social status vs. location sharing likelihood

### 3.3.6 Do The Users Have Specific Temporal Biases While Sharing Location?

In this section we try to discover the temporal biases of the users while they share location. We define the weekend location sharing likelihood,

$p_{we} = \frac{Number\, of\, geo-located\, Tweets\, on\, weekends}{Total\, number\, of\, Tweets\, on\, weekends}$. Similarly, we define weekday location sharing likelihood

$p_{wd} = \frac{Number\, of\, geo-located\, Tweets\, on\, weekdays}{Total\, number\, of\, Tweets\, on\, weekends}$ We have used the UTC time and UTC offset field to calculate the local time of the Tweets. We plot the users as circles on the axes representing $p_{we}$ and $p_{wd}$ in Figure 3.15. Similarly we calculate the morning $p_m$(6am - 12am), afternoon $p_a$ (12am - 6pm) and night $p_n$ (6pm-3am) location sharing likelihoods of all the users and plot them in Figures 3.16, 3.17 and 3.18. We now discuss these plots and make observations in the following part of this section.

In Figure 3.15 the x-axis represents the weekend location sharing likelihood and the y-axis represents the weekday location sharing likelihood, i.e. $p_{we}$ and $p_{wd}$ respectively.

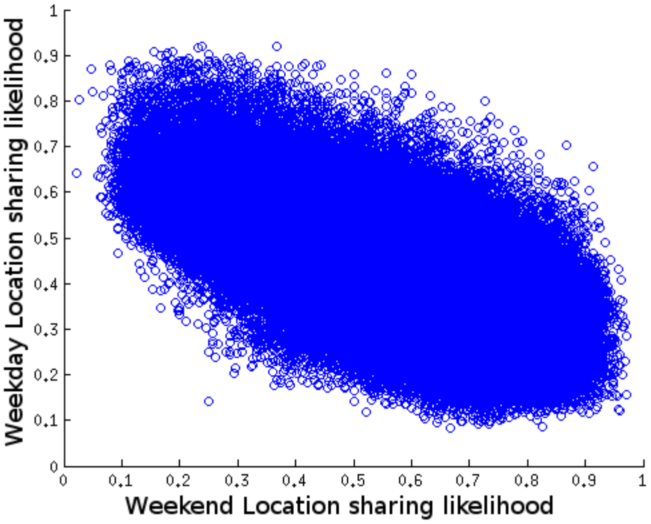

Figure 3.15: Weekend vs. weekday location sharing likelihood.

In Figure 3.16 the x-axis represents the morning location sharing likelihood and the y-axis represents the afternoon location sharing likelihood, i.e. $p_m$ and $p_a$ respectively.

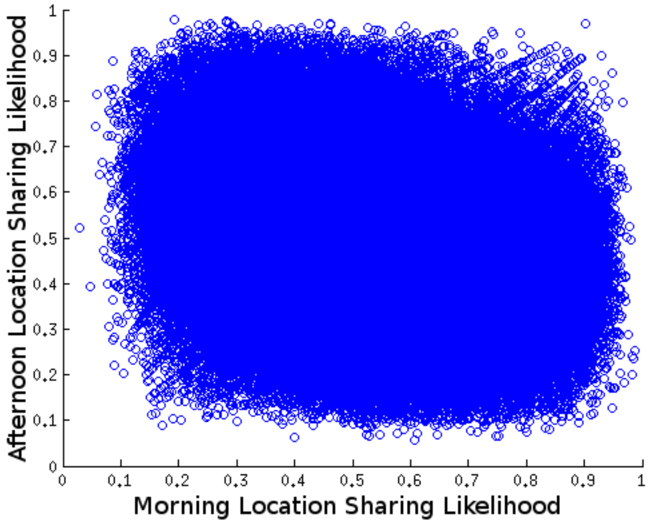

Figure 3.16: Morning vs. afternoon location sharing likelihood.

In Figure 3.17 the x-axis represents the morning location sharing likelihood and the y-axis represents the night location sharing likelihood, i.e. $p_m$ and $p_n$ respectively.

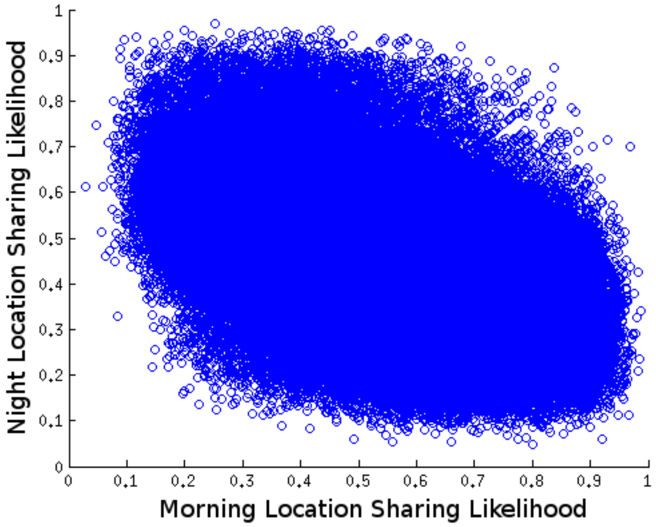

Figure 3.17: Morning vs. night location sharing likelihood.

In Figure 3.18 the x-axis represents the afternoon location sharing likelihood and the y-axis represents the night location sharing likelihood, i.e. $p_a$ and $p_n$ respectively.

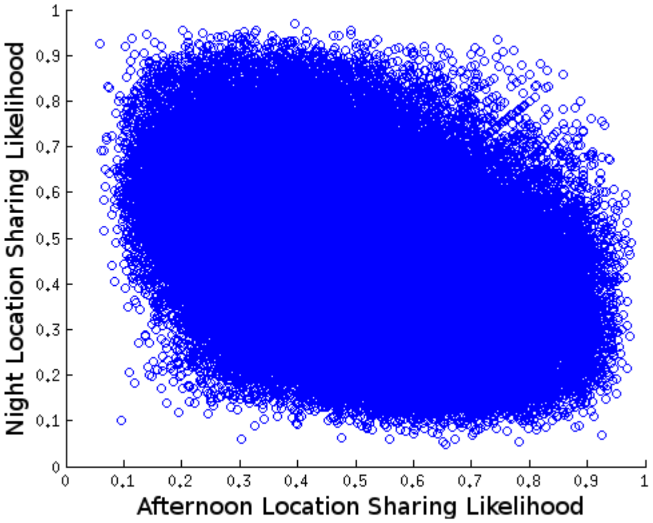

Figure 3.18: Afternoon vs. night location sharing likelihood.

### *3.3.7 Observations*

- In Figure 3.15 we see that users are have different biases, some users are more likely to share location during the weekends while others are more likely to share on the weekdays. Also, the more dense region in the lower right of the plot indicates that more users are biased to share location on weekdays than on weekends.

- We do not see any users on the lower left and upper right corners of the plot in Figure 3.15 unlike the plots in Figure 3.16, 3.17 and 3.18 which implies that more users are concerned about the *day of week* rather than the *time of day* when it comes to location sharing.

### *3.3.8 Is There a Correlation Between Tweet Source and Location Sharing?*

We plot the overall likelihood of each user using a mobile device vs the likelihood of sharing location.The Tweets with source field as *iPhone, Android, BlackBerry, iPad, BlackBerry, TweetDeck, Mobile, foursquare, iOS, Echofon or Phone* are considered to be mobile generated Tweets. The Mobile Source likelihood for each user is :

$$p_{mobile} = \frac{Number\, of\, mobile\, Tweets}{Total\, number\, of\, Tweets\, generated\, by\, the\, user}$$

Figure 3.19 shows the plot of the mobile source likelihood vs location sharing likelihood. The plot clearly show that there is a strong correlation between the location sharing likelihood and mobile source likelihood. Therefore, an increase in mobile usage makes it more likely that the users will share location.

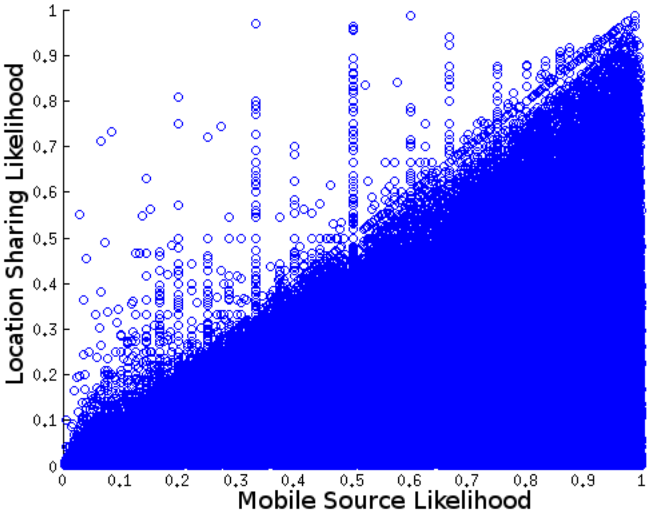

Figure 3.19: Mobile source likelihood vs. location sharing likelihood

## 3.4 Summary

In this chapter we presented an analysis of the location sharing behavior of social media users through a data-driven investigation of Twitter. We first gave a brief overview of the data and then we classified the users as All Sharers, Selective Sharers and Non Sharers. We analysed the content of these categories and came to know that the Selective Sharers have a language model close to that of the language model of spoken English. Then we further studied the location sharing behavior of Selective Sharers with respect to the content generated by them, their temporal activity, their social status and the devices they use for generating the Tweets. We classified the most informative CMU-ARK [26] word clusters for the geo-located Tweets under these different categories. We also found out that social status and mobile source were the two most correlated features to location sharing likelihood. We saw in Section 3.3.6 that the temporal biases of the users show up in the weekend vs weekday plot while they are not so prominent in the daily activity plots.In the next section we build predictive models for the location sharing behavior of the Selective Sharers.

# 4. MODELS

In the previous chapter we observed how certain features of Tweets generated by users and their profiles correlate to their location sharing behaviour. In this chapter we model the location sharing behaviour of users toward identifying common factors which influence location sharing in social media and also to evaluate their relative importance in predicting location sharing behavior. We achieve these objectives by modeling the location sharing behavior using a global model in Section 4.1. Later on, we proceed to build individual models in Section 4.2, to develop an insight into the predictability of location sharing behavior of Twitter users, because it can help us to have an idea about their vulnerability to potential location privacy attacks. We discuss our findings in section 4.3 with some remarks on the scope of the applicability of these results.

## 4.1 Global Model

Firstly, we try to model all the users with a single global classifier in an attempt to find out what are the different attributes of users and their Tweets which play the most important role in predicting their location sharing behavior. We briefly discuss the various features that we use for the model then we discuss the classifier that we use to model location sharing behaviour. We broadly use two kinds of features for our model:

1. Tweet Features: These are the attributes of the Tweets generated by the user and are therefore Tweet specific. The following are further categories of features that we use for our model:

(a) Temporal Features: These features include the hour of day and the day of week when the Tweet was generated.

(b) Contextual Features: We label the words in the Tweet using the LIWC[28] labels and then we represent each Tweet as a vector of length 64 where each dimension of the vector corresponds to an LIWC category. More formally, we have a Tweet, $\mathbf{T} = \{w_1, w_2, ..w_n\}$, where $w_i$ represents a word in the Tweet. For each $w_i$ we find category $c_j$ in the set of LIWC categories $C$, we ignore all the words which could not be mapped to the LIWC categories. Now, we have vector $\mathbf{V}$ in which each component $v_j$ represents the number of times a word with category $c_j$ was encountered in the Tweet $\mathbf{T}$.

(c) Additional Features: The other features include *boolean features* like whether the Tweet has a mention in it or not, whether it has a hashtag or not, source of the Tweet (mobile or other),

2. User Profile Features: We use social status defined in section 2.2.3 as the user profile feature. This feature was found to be correlated with the location sharing behaviour of the users and therefore we include it in our model.

We train a decision tree classifier available in the scipy API [30] over the dataset containing 32K Tweets with 50% Tweets belonging to each category i.e. geo and non geo respectively. We then test the classifier on the datasets with 80K Tweets having 20% geo Tweets and 80% non geo Tweets. We repeat this setup for three such test and training samples drawn from the Selective Sharers dataset. We measure the classifier using averages of several standard metrics.

We find an average accuracy of 0.68 and an average ROC (Receiver Operating Characteristic) area under the curve score of 0.56. These scores indicate that the global classifier did not perform very well maybe because the Twitter users comprises of a diverse group of users and although they have some common factors influencing their location sharing behavior, they also have significantly different behavior. We try to investigate this hypothesis in the next section. Based on this global classifier, we further identified the most significant features by measuring their importance scores. The importance score of a feature is computed as the (normalized) total reduction of the criterion brought by that feature. It is also known as the Gini importance. Following are the features in the order of importance scores generated by the decision tree classifier:

$$source\,of\,Tweet > social\,status > mention > leisure > body > bio > i > friend >$$

$$quant > health > feel > excl > you > shehe > percept > past > funct.$$

The most important features are the social status of the user and the source of the Tweet, i.e. mobile vs web. Then Twitter mentions followed by the LIWC [29] categories comprise the most important features. So, overall the social factor plays an important role in the location sharing behaviour of the users besides the device used by the user for generating the Tweet.

## 4.2 Individual Model

As we see, the performance metrics obtained from our model obtained are not very good, since we train the classifier over all the Tweets from all the users. And arguably not all the users have the same location sharing behaviour. Probably that is why the global model does not perform so well. Following this intuition we build individual models, that is, we train a different model for each user and then we report the obtained metrics for these individual models similar to what we did for global classifier. The rest of this section explains our methodology and the obtained results. We randomly sample 4700 users from the dataset with minimum number of 50 Tweets and with a minimum of 5 Tweets in geo-located and non-geo-located category. We do this in order to deal with data sparsity. For each user, we obtain a training set, which is generated from 70% of the total Tweets for the user, having 50 % Tweets belonging to the geo-located and non geo-located category, and a testing set which is generated from the remaining 30% has the actual proportion of the geo-located and nongeo-located Tweets. Figure 4.1 shows the scores for the 4700 users for whom we train our model. We see that more 60% of the users have an ROC score above 0.95 and accuracy score above 0.9. Therefore, our model is able to represent the user location sharing behaviour efficiently when we train it for an individual user.

### 4.3 Discussion

The experiments in the previous sections show that while users have common motivations and deterrences impacting their location sharing behavior, they have different levels of predictability. Although, we did not investigate into the reasons for this but it could be that some users are more diverse in nature, for example, they travel a lot, talk about different topics or places. In section 4.1.1, we also see that the LIWC categories like body, i, appear at the top followed by friend, feel, excl, you, she, he. While the former group of LIWC categories represents a sense of individuality, the later group suggests that the social factors are also the driving factors for location sharing. Our results, of course, hold only for Twitter as a social media platform and may or may not be generalizable to social media. Twitter is unique in the sense that it is a public social media platform, where there is almost no level of discretion while posting content. Therefore, users are naturally more hesitant to share location or more intimate details about themselves. We surmise that it is one of the reasons

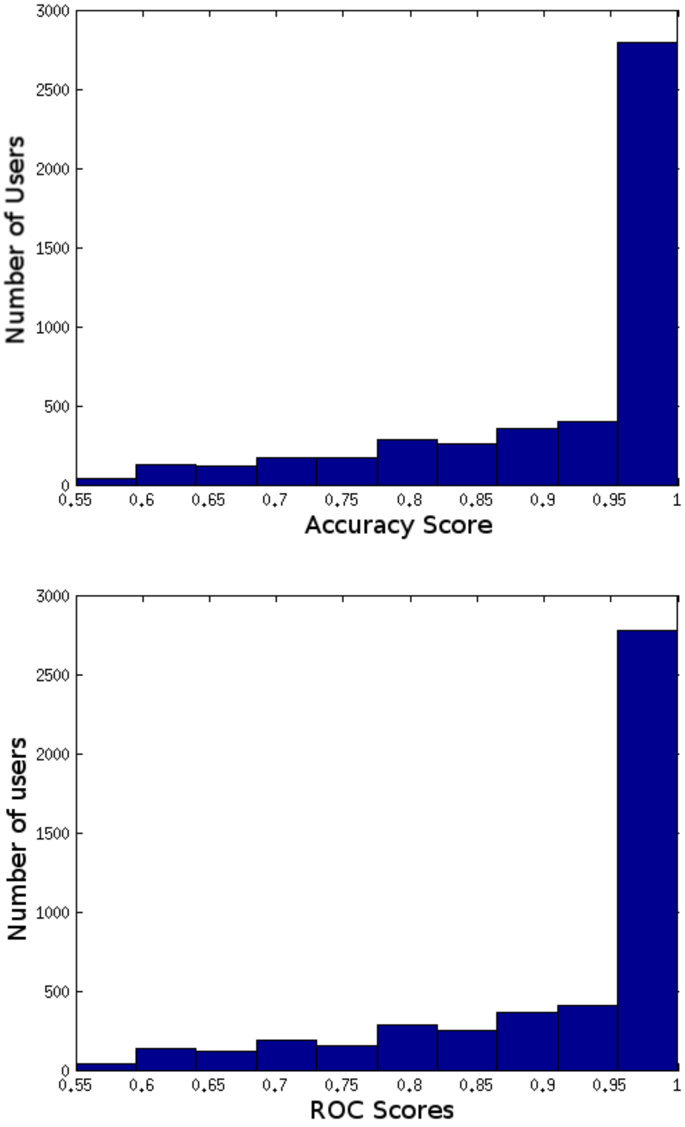

Figure 4.1: Accuracy and ROC scores for the modelled users.

that *social status* acts as one of the most important features influencing the location sharing behavior of Twitter users. So, our results may be biased in that sense. The same kind of analysis for other platforms like Facebook or Foursquare may reveal somewhat different aspects of location sharing.

# 5. CONCLUSION AND FUTURE WORK

In the previous section we modelled the location sharing behaviour of the users. Firstly, we try to model all the users using a global model. We train the global model using all the tweets from the selective users. As expected, we get a low performance with this model. Then we build models for individual users and we see that we are able to model more than 60% of the users with more than 90% accuracy and 0.9 or above ROC score as shown in the previous section. We infer the following from these results:

- users have very different location sharing behaviours and therefore global model does not model their behaviour efficiently and hence we need individual models for achieving a reasonable performance.
- most of the users have a highly predictable location sharing behaviour.
- *social status* and *source of the tweet* (mobile vs non mobile) are the most important features for the global model.
- the *mention* feature shows that there is an associated social context with geo-location, i.e. people mention other people while sharing location.

Our individual location sharing model is suitable for :

- developing location privacy protection mechanism based on this data driven model of the location sharing in the context of social media.
- developing an application programming interface which can be used by various location based services to provide push notifications to the user based on the

users' individual location sharing model. The decision tree model essentially represents users' mental framework and therefore it can be used to develop such services which can customize their behaviour based on this model.

The location sharing behavior model can be improved as follows:

- Our model does not use the venue information which is also an important factor which influences the location sharing behavior of the user. Therefore, it can be improved by using venue information from sources like Foursquare, Openstreetmap or Facebook.

- Currently, our model uses the CART decision tree which can be replaced by ID3 decision trees which are suitable for online learning and such a model can adapt to the changing location sharing behaviour of the users.

# BIBLIOGRAPHY

[1] Kelley, P. G., Benisch, M., Cranor, L. F., & Sadeh, N. "When are users comfortable sharing locations with advertisers?." *Proceedings of the Special Interest Group on Computer-Human Interaction Conference on Human Factors in Computing Systems.*. ACM, 2011.

[2] Lin, Jialiu, et al. "Expectation and purpose: understanding users' mental models of mobile app privacy through crowdsourcing." *Proceedings of the 2012 ACM Conference on Ubiquitous Computing.* ACM, 2012.

[3] Lin, J., Amini, S., Hong, J. I., Sadeh, N., Lindqvist, J., & Zhang, J. "A classification of location privacy attacks and approaches." *Personal and Ubiquitous Computing*18.1 (2014): 163-175.

[4] Toch, Eran, Justin Cranshaw, Paul Hankes Drielsma, Janice Y. Tsai, Patrick Gage Kelley, James Springfield, Lorrie Cranor, Jason Hong, and Norman Sadeh N. "Empirical models of privacy in location sharing." *Proceedings of the 12th ACM International Conference on Ubiquitous Computing.*ACM, 2010.

[5] Benisch, M., Kelley, P. G., Sadeh, N., & Cranor, L. F. "Capturing location-privacy preferences: quantifying accuracy and user-burden tradeoffs." *Personal and Ubiquitous Computing*15.7 (2011): 679-694.

[6] Barkhuus, Louise, and Anind K. Dey. "Location-Based Services for Mobile Telephony: a Study of Users' Privacy Concerns." *International Conference on Human-Computer Interaction - INTERACT.* Vol. 3. 2003.

[7] Tsai, J. Y., Kelley, P., Drielsma, P., Cranor, L. F., Hong, J., & Sadeh, N. "Who's viewed you?: the impact of feedback in a mobile location-sharing application."

*Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, 2009.

[8] Bigwood, Greg, Fehmi Ben Abdesslem, and Tristan Henderson. “Predicting location-sharing privacy preferences in social network applications.” *Proceedings of AwareCast* (2012).

[9] Wilson, Shomir, Justin Cranshaw, Norman Sadeh, Alessandro Acquisti, Lorrie Faith Cranor, Jay Springfield, Sae Young Jeong, and Arun Balasubramanian “Privacy manipulation and acclimation in a location sharing application.” *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing.* ACM, 2013.

[10] Wagner, D., Lopez, M., Doria, A., Pavlyshak, I., Kostakos, V., Oakley, I., & Spiliotopoulos, T. “Hide and seek: location sharing practices with social media.” *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services.* ACM, 2010.

[11] Cramer, Henriette, Mattias Rost, and Lars Erik Holmquist. “Performing a check-in: emerging practices, norms and 'conflicts' in location-sharing using foursquare.” *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services.* ACM, 2011.

[12] Lindqvist, J., Cranshaw, J., Wiese, J., Hong, J., & Zimmerman, J. “I'm the mayor of my house: examining why people use foursquare-a social-driven location sharing application.” *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, 2011.

[13] Page, Xinru, Alfred Kobsa, and Bart P. Knijnenburg. “Don't disturb my circles! Boundary preservation is at the center of location-sharing concerns.” *Interna-*

*tional Conference on Web and Social Media.* 2012

[14] Oh, Yuna, Kangsoo Jung, and Seog Park. "A privacy preserving technique to prevent sensitive behavior exposure in semantic location-based service." *Procedia Computer Science* 35 (2014): 318-327.

[15] Zhao, Yuchen, Juan Ye, and Tristan Henderson. "Recommending location privacy preferences in Ubiquitous Computing." *Conference on Security and Privacy in Wireless and Mobile Networks (WiSec)* . ACM, 2014

[16] Xie, Jierui, Bart Piet Knijnenburg, and Hongxia Jin. "Location sharing privacy preference: analysis and personalized recommendation." *Proceedings of the 19th International Conference on Intelligent User Interfaces.* ACM, 2014.

[17] Krumm, John. "A survey of computational location privacy." *Personal and Ubiquitous Computing* 13.6 (2009): 391-399.

[18] Minch, Robert P. "Privacy issues in location-aware mobile devices." System Sciences, 2004. *Proceedings of the 37th Annual Hawaii International Conference on. IEEE*, 2004.

[19] Tang, K. P., Lin, J., Hong, J. I., Siewiorek, D. P., & Sadeh, N. "Rethinking location sharing: exploring the implications of social-driven vs. purpose-driven location sharing." *Proceedings of the 12th ACM International Conference on Ubiquitous Computing.* ACM, 2010.

[20] Consolvo, S., Smith, I. E., Matthews, T., LaMarca, A., Tabert, J., & Powledge, P. "Location disclosure to social relations: why, when, what people want to share." *Proceedings of the SIGCHI Conference on Human factors In Computing Systems.* ACM, 2005.

[21] Lee, B., Oh, J., Yu, H., & Kim, J. “Protecting location privacy using location semantics.” *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 2011.

[22] Patil, S., Norcie, G., Kapadia, A., & Lee, A. J. “Reasons, rewards, regrets: privacy considerations in location sharing as an interactive practice.” *Proceedings of the Eighth Symposium on Usable Privacy and Security.* ACM, 2012.

[23] N. Sadeh, J. Hong, L. Cranor, I. Fette, P. Kelley, M. Prabaker, and J. Rao, “Understanding and capturing people's privacy policies in a people finder application, *The Journal of Personal and Ubiquitous Computing*, vol.13, Aug. 2009, pp. 401-412.

[24] Cheng, Z., Caverlee, J., Lee, K., & Sui, D. Z. “Exploring millions of footprints in location sharing services.” *International Conference on Web and Social Media 2011* (2011): 81-88.

[25] “The Twitter Rules” : https://support.twitter.com/articles/18311-the-twitter-rules *(Accessed on 20 April, 2015)*

[26] Owoputi, O., O’Connor, B., Dyer, C., Gimpel, K., Schneider, N., & Smith, N. A. ”Improved part-of-speech tagging for online conversational text with word clusters.” *North American Chapter of the Association for Computational Linguistics Human Language Technologies 2013.*

[27] “CMU ARK Homepage” : http://www.ark.cs.cmu.edu/ *(Accessed on 20 April, 2015)*

[28] Pennebaker, James W., Martha E. Francis, and Roger J. Booth. ”Linguistic inquiry and word count: LIWC 2001.” *Mahway: Lawrence Erlbaum Associates.* 71 (2001): 2001.

[29] "LIWC Categories Description": http://www.liwc.net/descriptiontable1.php *(Accessed on 20 April, 2015)*

[30] "Scipy API" : http://www.scipy.org/ *(Accessed on 20 April, 2015)*

[31] Li, Nan, and Guanling Chen. "Sharing location in online social networks." *Network, IEEE* 24.5 (2010): 20-25.

[32] "The Twitter Sampling API reference documentation": https://dev.twitter.com/streaming/reference/get/statuses/sample *(Accessed on 20 April, 2015)*

[33] "The Infolab Home Page": http://infolab.tamu.edu *(Accessed on 20 April, 2015)*