An Exploration of Social Interaction on Twitter
Bethany Waldmann*, Jeremy Kelley, and James Caverlee
Texas A&M University

Abstract—With the rapid rise in the past few years of large-scale social media (e.g., blogs, Facebook, YouTube), the Web is fundamentally transforming into a Social Web centered around users and their connections to other users. In this project, we have studied the geographic connections among Social Web users by analyzing Twitter, one of the most buzz-worthy recent Social Web successes. Twitter is a microblogging service that has attracted millions of users who communicate via short messages of 140 characters or less (“tweets”). Concretely, we have explored the geographic connections among Twitter users through the application of spatial interaction theory. Spatial interaction theory is one of the cornerstones of geographic theory, and is used to model and understand the linkages and flows between locations. In this project, we have: (i) crawled and extracted location-based information from approximately 1.02 million Twitter users; (ii) validated this location information using MetaCarta, an online service which returns all possible location matches for a given input; (iii) constructed a Twitter communication graph by mining @mentions from tweets (placing the ‘@’ symbol before a username in a tweet forms a virtual link to another user; viewing these links in the aggregate yields the Twitter communication graph); and (iv) analyzed the location distribution of Twitter users and their inter-location communication patterns. We find that a majority of users are from the United States. Comparing interaction among countries and Metropolitan areas in the US, we infer that physical distance has some affect across the scales but is somewhat irrelevant when considering the scales separately.

Index Terms—Communication system, geographic information systems, modeling, networks, social media, social web.

1 INTRODUCTION

Recent innovation has diversified use of the Web by transforming it from the one-way transfer of information from tech-savvy website owner to less-savvy Internet surfer it has traditionally been in the past decade or two by adding to it a means for the general public to communicate with each other. This dissemination of personal information is present through the spectrums of online dating, blogs, MySpace, Facebook, and now Twitter, each having one or more aspects with increased public availability over traditional methods such as face-to-face conversation, letters, email, and instant messaging.

1.1 Twitter

Twitter is a social network in which members write short messages about what they are doing, who they are with, a topic of interest, or anything else they wish to express. These short messages, also called “tweets”, are limited to 140 characters or less and are completely public unless the member takes initiative and purposefully changes a privacy setting. Although the majority of messages are public, a member...
can “follow” other users to specify that they are interested in the users’ tweets; this results in the most recent messages from all followees to be displayed on the home feed of the follower.

Although Twitter does not release site statistics, a number of reports estimate that there are approximately 7 million accounts to date. Tweets are indexed chronologically as integers and number over three billion, so with a current rate of 600,000 to 850,000 tweets per day the count is quickly approaching the maximum unsigned 32-bit integer: 4,294,967,295. This is an immense amount of volunteered information covering a wide range of topics from personal information and life stories, to local and world news, to pure facts and other material of possible interest.

In short, Twitter is acting as an extremely large human sensor network which scientists and other curious people can mine for user-given information. Many sites have picked up the Twitter trend and are creating mashups, which are an aggregation of information from various sources in a single place, but there has been little academic research done into what is available and how it can be used. Various applications involving geographic information supplied could include, among other things: finding trends in sickness before people are ill enough to see a doctor, who will only notice the progression after it has escalated; creating maps of topical interest; and pulling tweets in real-time from disaster areas where live coverage, by media or aiding organizations, is not feasible.

### 1.2 Spatial Interaction

Spatial interaction describes flow, or movement, of some element among locations. The element could be anything, ranging from people commuting, to materials being shipped, to ideas being exchanged. To model this, geographers developed a theory that specifies three causes for the flow of an element: complementarity, transferability, and intervening opportunity [1]. Complementarity is the cause for the flow; a lack in one place and an abundance in another is plenty to motivate, for example, the export of oranges from tropical Florida to arctic Alaska. Transferability determines the tenability of an element to flow between two given locations; transportation of oranges is constrained by distance, time, and shipping infrastructure, whereas conveyance of information is more often hindered by inavailability of technological infrastructure. Intervening opportunity gives higher probability for interaction to occur between closer locations when there are multiple sources or sinks for the element to flow from or to; if oranges are grown in sunny California, which is much closer to Alaska than Florida, it is much more logical for Alaska to buy their oranges from California since the distance and time of transportation will be reduced, decreasing cost and yielding fresher fruit.

### 1.3 Goals

With such a large amount of information available there are many research possibilities. Before undertaking any large-scale problem it is necessary to consider some basic questions about Twitter. We will consider spatial interaction among users, which requires three fundamental research questions (RQs):

#### 1.3.1 RQ1: What location information is available?

Judicious use of data is extremely important, so rather than immediately delving into expansive research on voluminous data it is prudent to consider a smaller set to diminish time wasted processing ill-represented data. We are specifically looking to see various ways users are embedding information about where they are, as well as how prevalent it is over the set of users.

#### 1.3.2 RQ2: How are users distributed?

After determining the best way to ascertain users’ locations, we will be able to commence analysis on a larger set of data. In particular we are interested in where a majority (if one exists) of users are from. Are most users from the United States? How well are countries represented? Is it feasible to consider interaction

---

among countries, or are users aggregated in one country? Would it be possible to consider interaction among cities, or is the data too sparse?

1.3.3 RQ3: Is location affecting interaction?

Examining interaction at different levels, we will analyze how location has an impact on interaction, if it does at all. Does distance matter, or is it irrelevant? Are users in some locations more prone to interact than users in other locations? Do factors of location other than distance have an effect on interaction?

2 Methodology

This section will describe the more technical details of how results were achieved.

2.1 Datasets

Our research questions require two datasets. These datasets are different sizes, contain varying elements, and cover different time periods.

2.1.1 Dataset A

To test available information, we consider a reasonable dataset of 1000 users’ complete history, spanning from account creation through April ’09. It was compiled by another student prior to this research by scraping Twitter’s public timeline3 and accessing their large public API feed4.

2.1.2 Dataset B

For analyzing distribution and interaction, we consider a sizable dataset of 1,020,062 users; although this is not the full collection of Twitter users, it represents approximately 15% of its population and contains @ mention metadata (defined later). Another student compiled this set which covers a 10-month interval, April ’08 through January ’09, and was obtained by crawling Twitter.
2.2 Analyzing Available Information

Here we utilize Dataset A to find out what options there are for mining a user’s location. There are some options one has to indicate a location, and the likely obvious is to specify a location in the user profile. However, people are bound to move around without changing that location because not only would it be rather annoying, especially if the user moves around a lot, but the location in the profile is generally accepted as the home location. The simple solution: put it in a tweet! Still, there are multiple ways a user can do this. The first is the most intuitive: simply put the place name in the tweet, such as “On vacation in Rome!” A second way is to use hashtags, in which you put a hash mark, or pound or number sign, in front of a word to indicate that the tweet belongs in the following category; to modify the previous example, it could look like “On vacation! #Rome”, “In Rome! #vacation” or “On #vacation in #Rome” depending on the user’s preference. A third way is to use an “L:” before the location; again modifying our example, “On vacation in L:Rome!” Finally, airport codes are often used; to adapt our example further, “Just landed at CIA - vacation has begun!”

2.3 Mining Locations

Dataset B is used for the remaining research. Since the user’s profile is not included in this dataset, locations are determined by crawling the web. After obtaining the text a user lists as their location, it is necessary to correlate it with a specific location on the earth; if the text is “Boston,” is the person from Boston, Massachusetts, or Boston, United Kingdom, or someplace else? To answer this question we use an online geocoding service, MetaCarta, which inputs a string of text and outputs all possible locations this text could be referring to; output for each location includes, among other things, a path (e.g. Boston, Suffolk, Massachusetts, United States), coordinates specifying the center of the area, and the population. Locations are listed in descending order of population (i.e. the most probable to the least probable). Although the most probable may not be the user’s actual location, there is no way to determine the correct location without asking the user. For the Boston example, MetaCarta lists 69 possibilities, the top one being Boston, Massachusetts.

Since MetaCarta is not human it has a lesser understanding of its input; some of the results of this are: syntax constrained to areas of increasing size separated by commas (e.g. Boston, Massachusetts, United States); inability to process coordinates; and intolerance for misspelling, nicknames, and many abbreviations.

2.4 Defining a Communication Network

Although “following” defines links between users, it is extremely formal and does not indicate any real relationship; it simply implies that one user is interested in what another user is saying. Instead, we consider the graph of @ mentions, which is a way of mentioning another user or specifying that a tweet is specifically for them, although it can still be seen by anybody; this is done by putting an “@” in front of their username. This also allows users who are not following each other to communicate, since all tweets mentioning a user is available to them (only) on a separate feed. For example, one might tweet “@THE_REAL_SHAQ you were really awesome in the game last
night!” to support Shaquille O’Neal without getting all of his tweets.

The @ mention metadata referenced above are a set of files containing, for each user, the timestamp and end user of each @ mention made during the sample period. The communication network is defined as the graph formed with users as nodes and directed edges of @ mentions between them. There is a maximum of two directed edges between any two given users (one each way), simply indicating whether each has @ mentioned the other.

2.5 Measuring Interaction

A few steps are needed to determine the interaction between any two given places. For each path $A$ found while retrieving users’ locations, we define $ResA$ as the set of all residents, or users listing $A$ as their location. For each path $A$ we also define $ConA$ as the set of all contacts, or users who have been @ mentioned by $ResA$. Finally, we define the interaction between two locations $A$ and $B$ as

$$\| ResA \cap ConB \| + \| ResB \cap ConA \| .$$

3 Results

In this section we determine the granularity at which to process our data based on the distribution of our sample and discuss the resulting maps of interaction.

3.1 Syntax Prevalence

Here we test the various syntax options for frequency of use and viability. Since the plain-body text and hashtag options would require a list of all locations to check words and phrases against, these were skipped; even if such a list existed or was compiled by us, use would be limited because of the time it would take to process all the data. A quick test showed that 66% of the sample set listed a location in their profile. Moving on to analyze content of the tweets themselves, we found that only 4% used the “L:” syntax and 90% used airport codes.

Considering Rome’s code (CIA), the extreme success of airport codes was not surprising; however, analysis of the tweets containing these codes was deemed necessary to validate our hypothesis that a majority of references were due to common acronyms and overzealous capitalization. This closer inspection, although not covering all 83,183 references, confirmed our theory. It also stimulated the possibility to form a set of words that, when included in the tweet, would indicate a much higher likelihood that the code was indeed referring to an airport; a set like this might include words such as “airport,” “plane,” and “fly.” However, since this would require a much more intense study of airport language than the simple analysis planned, we leave this to future research.

The conclusion of this preliminary research answering RQ1 is that the most facile and effective indication of a user’s permanent location is what is listed in the user profile.

3.2 Location Distribution

Approximately 9% of the locations had no matches in MetaCarta. A majority of the remaining population is from the United States, but there are a number of countries with knowledgeable representation; it is plausible to consider country-to-country interaction. Countries that were picked had over 1000 users who represented at least .01% of their actual population.

Since the United States has the largest representation by far, it was chosen for analyzing interaction on a smaller level. An initial test
of city-to-city interaction in Texas yielded poor results, so the top 25 Metropolitan Statistical Areas (MSAs), which cover multiple counties each, were opted for instead.

Thus we answer RQ2: city population as found by our parsing implementation on a fractional dataset is too sparse, so we will consider interaction at two granularities: MSAs and countries.

3.2.1 Error

Iran is shown to have the 4th largest population; this is a misrepresentation of geographic distribution. Due to political events surrounding the recent election in Iran, residents were using Twitter to organize protests; the authorities caught on and started using Twitter’s location search to prosecute these residents. As a result Iranians asked people to change location to Iran, which would generate false positives and make it harder for the authorities to locate actual residents.

This draws attention to the fact that the information is user-provided and not necessarily accurate. Some users will list imaginary places such as “Twitterland”; some will pretend to be from novel places like Timbuktu (Africa), Worlds End (Massachusetts), or The Bar (Australia); and some will falsify their location by substituting Colorado for Alabama. Although this is bound to happen, there is relatively little motivation to do so and can be circumvented fairly well by disincluding the novel locations and being aware of phenomena such as that produced by the Iran election.

Despite the skewing of distribution that results, there is a benefit; namely, a measure of user support for a cause.

3.3 Spatial Interaction

MSA interaction is predominantly among MSAs on the east and west coasts, with Chicago the only large focus of communication in the central and mid-west regions. Given the vast amount of interaction among the MSAs, a higher threshold was required to keep the map readable; only 9% of the connections $^{5}$ and 42% of the relationships $^{6}$ are shown.

Interaction with and among countries having a primary language of English seems to be prevalent. As such, the US, the UK, Canada, and Australia serve as hubs of communication. Although there is a considerable amount of interaction between countries, it does not compare to the interaction within the US; however, the representation of individual US cities is much greater than those of individual countries, including the United States. In contrast to the MSA map, 16% of the connections and 91% of the relationships are shown.

Considering interaction at both granularities we can answer RQ3. It appears that distance is

---

5. Connections are between areas.
6. Relationships are between users.
Fig. 6. Spatial interaction among top Metropolitan Statistical Areas (MSAs) in the United States. Map obtained from Wikimedia Commons.

Fig. 7. Spatial interaction among top countries of the world. Maps obtained from Wikimedia Commons.
not dead as some would claim it to be. However, its role has been diminished, as is clearly shown by some of the heaviest communication traffic between the coasts at the MSA level and between countries separated by vast oceans.

4 Conclusion

We have determined that there are three ways users embed their location that are viable for mining: providing a location in the user profile; using the “L:” prefix in a tweet; and referencing an airport code in a tweet. The profile is the best indication of a user’s location for the purpose of examining the affect of location on interaction due to its permanence and popularity. Using this location, we find that the distribution of Twitter users is poorly correlated with actual populations; the majority of users are from the United States. Additionally, since location is user-provided it is not necessarily accurate and can fall prey to phenomena such as the Iran election.

Comparing interaction at the MSA and country levels we discern that, although location does have an effect on the amount of interaction between two places, the largest streams of communication are going on among coastal cities of the US at a finer granularity and among numerous continents at a larger granularity. At this higher level, it is unsurprising that language seems to be a stronger predictor of interaction than physical distance.

Now that some elementary location-based analysis of Twitter has been done, there are numerous possibilities for future work. To extend this idea of interaction various other aspects of location, including politics, economy, and religious beliefs, could be considered; this could allow an expansive gravitational model to be formed. Other usage patterns, such as whether tweets are of a diurnal nature, could be discovered. Finally, text-based analysis of tweets could be used for practical applications such as the discovery of outbreaks and tracking of disaster situations.

Acknowledgments

The authors would like to thank the National Science Foundation for funding this research, as well as Texas A&M, the Department of Computer Science and Engineering, and REU coordinator Theresa Roberts.

References