Exploring Millions of Footprints in Location Sharing Services

Zhiyuan Cheng, James Caverlee, Kyumin Lee
Texas A&M University
College Station, TX 77843
{zcheng, caverlee, kyumin}@cse.tamu.edu

Daniel Z. Sui
Ohio State University
Columbus, OH 43210
sui.10@osu.edu

Abstract

Location sharing services (LSS) like Foursquare, Gowalla, and Facebook Places support hundreds of millions of user-driven footprints (i.e., "checkins"). Those global-scale footprints provide a unique opportunity to study the social and temporal characteristics of how people use these services and to model patterns of human mobility, which are significant factors for the design of future mobile+location-based services, traffic forecasting, urban planning, as well as epidemiological models of disease spread. In this paper, we investigate 22 million checkins across 220,000 users and report a quantitative assessment of human mobility patterns by analyzing the spatial, temporal, social, and textual aspects associated with these footprints. We find that: (i) LSS users follow the "Lèvy Flight" mobility pattern and adopt periodic behaviors; (ii) While geographic and economic constraints affect mobility patterns, so does individual social status; and (iii) Content and sentiment-based analysis of posts associated with checkins can provide a rich source of context for better understanding how users engage with these services.

1 Introduction

In many ways analogous to the sensor systems embedded in the physical environment of planet earth, emerging real-time social systems are rapidly creating a web of social sensors that can potentially be used as sociometers to gauge diverse social indicators ranging from political views to consumer tastes to public opinions about key social issues to the mood of people at particular places and times. In practice, highly-dynamic real-time social systems like Twitter, Facebook, and Google Buzz have already published exabytes of real-time human sensor data in the form of status updates. Coupled with growing location sharing services like Foursquare, Gowalla, Facebook Places, and Google Latitude, we can see unprecedented access to the activities, actions, and trails of millions of people, with the promise of deeper and more insightful geospatial understanding of the emergent collective knowledge embedded in these activities and actions.

In terms of scale, the Foursquare service alone claims over 6 million registered users (Foursquare 2011) and nearly 1 million check-ins per day (Grove 2010). Like similar services, Foursquare allows users to "check in" at different venues (e.g., grocery stores, restaurants), write tips, and upload pictures and videos.[1] As in other online social networks, Foursquare users can make friends with each other, and monitor their friends' status and location. While users of Foursquare and related location sharing services may not be a representative cross-section of the whole human society, the data revealed through these services provides a fascinating and unique opportunity to study large-scale voluntarily contributed human mobility data, which could impact the design of future mobile+location-based services, traffic forecasting, urban planning, and models of disease spread.

Toward understanding the spatial, temporal, and social characteristics of how people use these services, we present in this paper a large-scale study of location sharing services. Concretely, we study the wheres and whens of over 22 million checkins across the globe. We study human mobility patterns revealed by these checkins and explore factors that influence this mobility, including social status, sentiment, and geographic constraints. To the best of our knowledge, this is the first effort to utilize location sharing services to study human mobility patterns and the corresponding factors which can affect mobility patterns.

2 Related Work

The role of geography and location in online social networks has recently attracted increasing attention. Facebook researchers analyzed the distance between Facebook users' social relations, and utilized locations of a user's friends' to predict the user's geographical location (Backstrom, Sun, and Marlow 2010). (Cheng, Caverlee, and Lee 2010) modeled the spatial distribution of words in Twitter's user-generated content to predict the user's location. Characterizing network properties in relation to local geography is studied in (Yardi and Boyd 2010). User behavior with regard to the location field in Twitter user profiles has been studied in (Hecht et al. 2011). (Lindqvist and others 2011) analyzed how and why people use location sharing services, and discussed the privacy issues related to location sharing services. Besides locations, researchers have also explored temporal dynamics associated with on-line social activities (Golder, Wilkinson, and Huberman 2007).

[1] According to the Wall Street Journal, "check in" was the 12th most popular word of 2010 (Cholera 2011)

Table 1: Distribution of Sources of Checkins

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

Figure 1: Global Distribution of Checkins

Analyzing and modeling mobility patterns has long attracted attention by researchers in fields like statistical physics, ubiquitous computing, and spatial data mining. For example, an analysis of 100,000 cellphone users' trajectories (Gonzalez, Hidalgo, and Barabasi 2008) showed that human mobility displayed simple reproducible patterns. The authors of (Brockmann, Hufnagel, and Geisel 2006) analyzed the circulation of bank notes in the US and concluded that human traveling behavior can be described mathematically on many spatio-temporal scales by a two parameter continuous time random walk model. A 93% potential predictability in user mobility was found across 50,000 cellphone users in (Song et al. 2010). (Zheng and others 2009) proposed a system to mine interesting locations and travel sequences from users' GPS trajectories. Researchers of (Humphries and others 2010) observed Lèvy Flight search patterns across 14 species of marine predators, with a few individuals switching between Lèvy Flight and Brownian motion as they traversed different habitat types.

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

3 Gathering Checkins

To begin our study, we first require a collection of checkins. Since personal checkin information on location sharing services like Foursquare, Gowalla, and Facebook Places is typically restricted to a user's immediate social circle (and hence unavailable for sampling) we take an approach in which we sample location sharing status updates from the public Twitter feed. Twitter status messages support the inclusion of geo-tags (latitude/longitude) as well as support third-party location sharing services like Foursquare and Gowalla (where users of these services opt-in to share their checkins on Twitter). We monitor Twitter's gardenhose streaming API ($\sim$1% of the entire Twitter public timeline), and retrieve users who post geo-tagged status updates. For each sampled user, we crawl up to a maximum of the most recent 2,000 geo-labeled tweets.

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

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

Filtering Noise: Many location sharing services provide some mechanism to verify that a user is actually at or near the venue where they are checking in (e.g., by cross-checking with a user's cellphone GPS) (Foursquare 2010), however, there can still be incidents of false checkins. Hence, we additionally filter out all checkins from users whose consecutive checkins imply a rate of speed faster than 1000 miles-per-hour (or faster than an airplane). In total, we filtered 294 users (0.1%) with sudden moves, yielding a final collection of 224,804 users and 22,388,315 checkins. More than 72% users have fewer than 100 checkins; 7.8% users have more than 300 checkins; and 3.6% users have more than 500.[2]

Locating Each User's "Home": Some of the analysis in the following sections requires that we first associate each user with a natural "home", so, for example, we can compare the properties of all users "from" New York City versus users "from" Los Angeles. Since users of location sharing services are not required to register a home location, we must algorithmically determine the home location. Note that choosing a user's home based on the center of mass of all

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

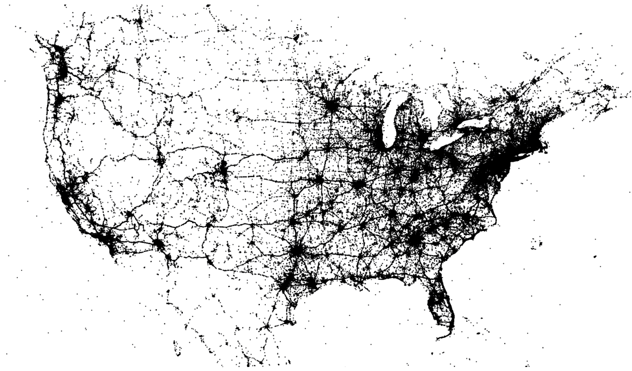

Figure 2: Detail: Checkins in the United States

checkins suffers from splitting-the-difference, by placing a user from Houston who occasionally travels to Dallas somewhere in between the two cities; alternatively, directly considering the user's most frequently checked-in venue may overlook a cluster of closely-located but less individually checked-in venues. To avoid these drawbacks, we propose a simple method to geo-locate a user's home based on a recursive grid search. First, we group checkins into squares of one degree latitude by one degree longitude (covering about 4,000 square miles). Next, we select the square containing the most checkins as the center, and select the eight neighboring squares to form a lattice. We divide the lattice into squares measuring 0.1 by 0.1 square degrees, and repeat the center and neighbor selection procedures. This process repeats until we arrive at squares of size 0.001 by 0.001 square degrees (covering about 0.004 square miles). Finally, we select the center of the square with the most checkins as the "home" of the user.

4 Spatio-Temporal Analysis of Checkins

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

4.1 Wheres of the Checkins

First, we plot the locations of the 22 million checkins in Figure 1, where we see that while checkins are globally distributed, the density of checkins is highest in North America, Western Europe, South Asia, and Pacific Asia. Zooming in on the US, Figure 2 shows the reach of location sharing services, revealing the boundaries of cities and the lines of highways. Further zooming in, we can see in Figure 3 how New York City is densely covered by more than $\frac{1}{2}$ million checkins. While these figures convey the scale and density of location sharing services, we can further explore the nature of these checkins by aggregating keywords across all 22 million checkin tuples. The aggregated view in Figure 4 shows that the most popular checkin venues are restaurants, coffee shops, stores, airports, and other venues reflecting daily activity (e.g., fitness, pubs, church).

4.2 Whens of the checkins

Considering the temporal distribution of checkins, we can uncover both the aggregate daily patterns of users of location sharing services and their weekly patterns. By normalizing the timestamps of every checkin so that all local times are treated as the same time (i.e., aggregating all checkins at 1pm, whether they be in Chicago or Tokyo), we show in Figure 5 the mean checkin pattern per day. This pattern provides a glimpse into the global daily "heartbeat", with three major peaks: one around 9am, one around 12pm, and one around 6pm. The diurnal pattern is clearly displayed as more people are active during the daytime than at night.

To illustrate the potential of location sharing services as sociometers of city health and activity, we show in Figure 6, the disaggregated daily checkin patterns of users in New York City, Los Angeles, and Amsterdam. The checkin patterns show that Amsterdam's daily "heartbeat" reflects an early-rising city, with more activity than either LA or New York in the morning hours. LA peaks around noon, whereas New York has the highest checkin rate during the night ("The City That Never Sleeps"). We are interested to further explore the reasons for these differences. Are the daily differences artifacts of local culture? Or the proclivity of users in certain locations to more willingly reveal certain aspects of their daily lives than others (e.g., checkin in while at work, but not at play?) Or do the differences reflect biases in the data, so that certain demographics are overrepresented in one city versus another?

Moving from the daily pattern to the weekly pattern, we

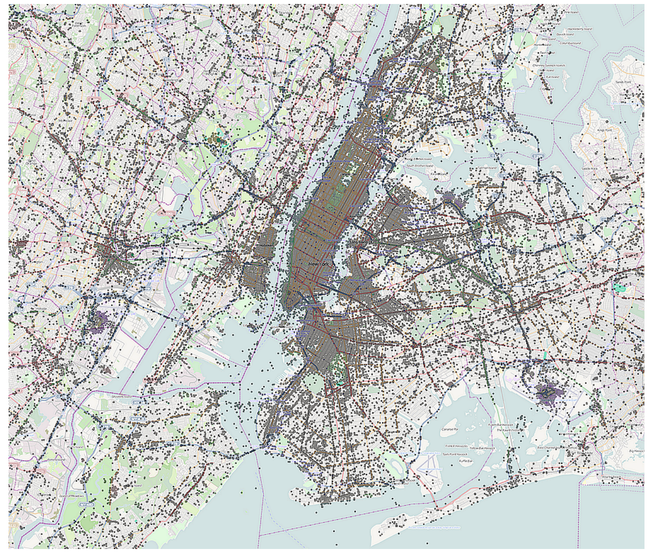

Figure 3: Detail: Checkins in New York City

Figure 4: Venue Cloud for Checkins

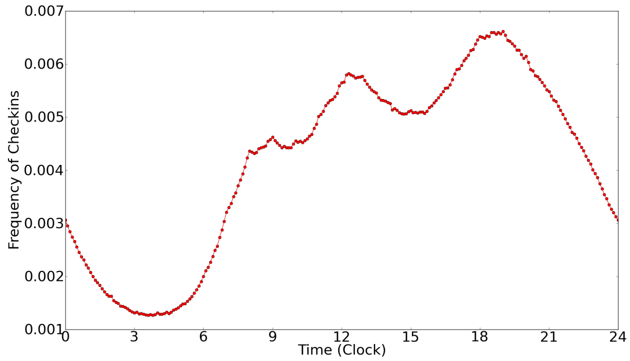

Figure 5: Mean Daily Checkin Pattern

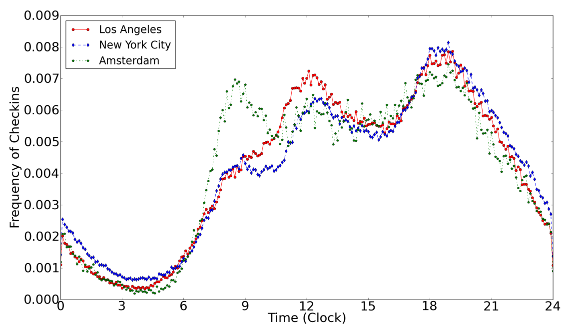

Figure 6: Daily Checkin Patterns: NYC, LA, Amsterdam

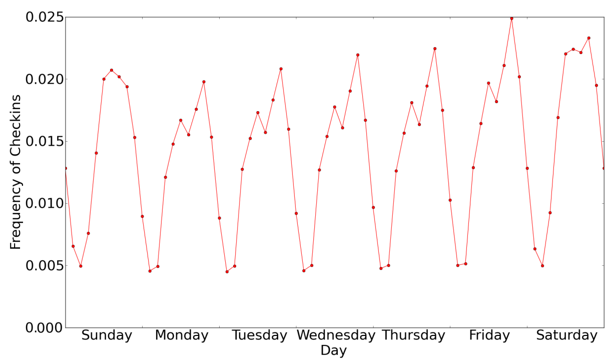

Figure 7: Mean Weekly Checkin Pattern

see in Figure 7 the aggregate global patterns over the days of the week. Weekdays clearly indicate two peaks during lunch time and dinner time, while over the weekend these two peaks blend, reflecting a fundamentally different weekend schedule for most users of location sharing services. We can also observe that the relative daily activity increases from Monday to Friday, peaking on Friday evening.

5 Studying Human Mobility Patterns

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

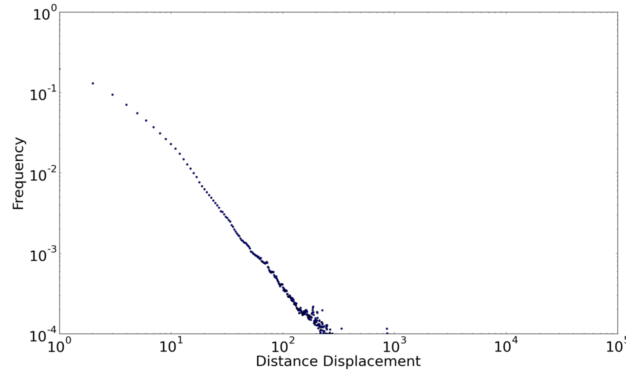

Figure 8: Distribution of Displacements

5.1 User Displacement

We begin with an investigation of the distance-based *displacement* of consecutive checkins made by users. Considering all pairs of consecutive checkins yields 22,163,511 separate displacements, reflecting the distance between these consecutive checkins (and hence, how far a user has traveled). We plot the distribution of displacement for the dataset on a log-log scale in Figure 8. The x-axis is the displacement in miles, and the y-axis is the frequency of displacements in the same bucket. The trend is approximated by a power-law:

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

where δ_r represents the displacement and $\beta = 1.8845$. The formula indicates that human motion modeled with checkin data follows a Lévy Flight (Rhee et al. 2008), in which a random walk proceeds according to steps drawn from a heavy-tailed distribution. A Lévy Flight is characterized by a mixture of short, random movements with occasional long jumps. Flight models with a similar scaling exponent have been observed separately in a study of displacements based on cellphone call data with $\beta = 1.75$ (Gonzalez, Hidalgo, and Barabasi 2008) and in a study of displacements based on bank note dispersal with $\beta = 1.59$ (Brockmann, Hufnagel, and Geisel 2006).

5.2 Radius of Gyration

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

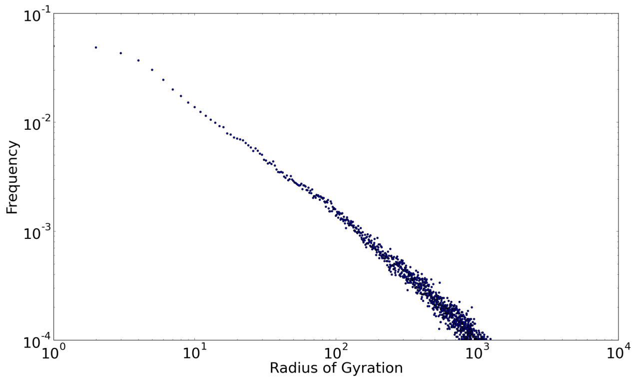

Figure 9: Distribution of Radius of Gyration

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

where n is the number of checkins of the user, and $(r_i - r_{cm})$ is the distance between a particular checkin r_i and the user's center of mass r_{cm} (which is a simple average location over all checkins). We calculate the radius of gyration for each user in our collection and the distribution of radius of gyration is displayed on A log-log scale in Figure 9. The x-axis identifies the radius of gyration in miles and the y-axis shows the number of users with that radius of gyration. The trend in Figure 9, like the distribution of displacements, also follows a power-law:

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

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

To illustrate how radius of gyration can give further insight into the dynamics of cities, Figure 10 plots the average radius of gyration of users in major cities (with 100,000+ population and at least 20 users in the checkin dataset) in the continental US. The red bubbles are cities with a radius of gyration larger than 500 miles; blue ones are cities with a radius larger than 250 miles; cyan ones have a radius larger than 125 miles, and yellow ones are the rest of major cities. Users in coastal cities tend to have a higher radius of gyration than users in inland cities, and people in central states tend to have a high radius of gyration due to long distance travels to the coasts. Even so, there are some interesting regional variations worth further study, for example, the low radius of gyration for El Paso compared to the higher radius for nearby Albuquerque.

5.3 Returning Probability

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

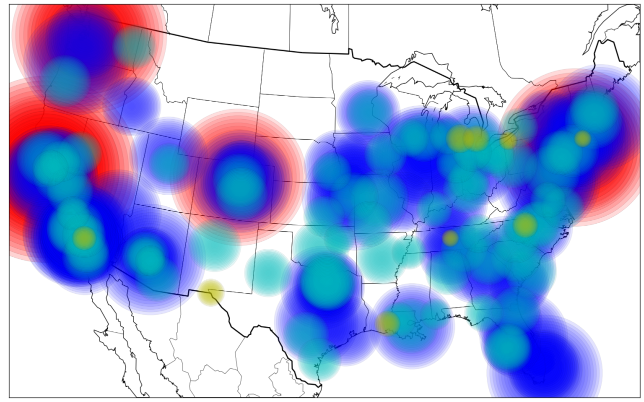

Figure 10: Mean Radius of Gyration for Users in US Cities

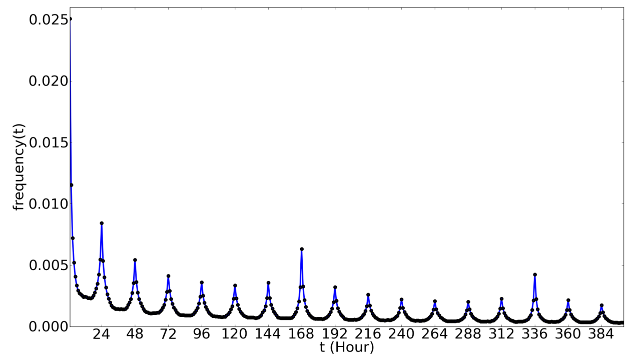

Figure 11: Distribution of Returning Probability

We measure periodic behavior by the *returning probability* (or, first passage time probability), which is the probability that a user returns to a location that she first visited t hours before. Grouping all returning times of all checkins into buckets of one-hour, we plot the distribution of returning times in Figure 11, in which the x-axis represents the bucket of returning time, and the y-axis is the corresponding frequency for a bucket. For example, at 168 hours, the returning probability peaks, indicating a strong weekly return probability. Similarly, we see daily return probabilities. As time moves forward, the returning probability shows a slight negative slope, indicating the aggregate forgetfulness of visiting previously visited places (that is, the return probability is strongest for places we have visited most recently).

6 Exploring Factors that Influence Mobility

In this final section, we turn our attention to exploring the factors that may impact human mobility. While factors like geography and economic status are natural to investigate, the unique properties of location sharing services provide an unprecedented opportunity to consider heretofore difficult to measure aspects of human behavior. For example, does social status as measured through popularity in these services impact a user's radius of gyration? Does user-generated con-

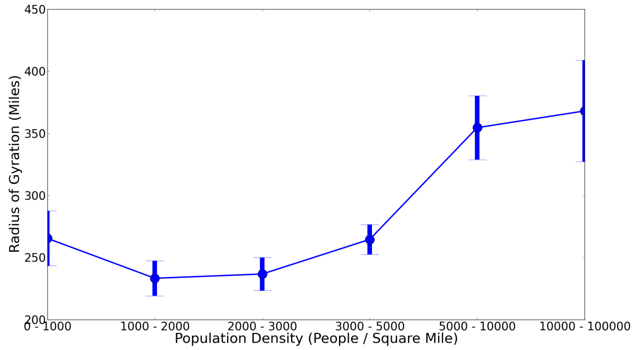

Figure 12: Average R_g versus City Population Density

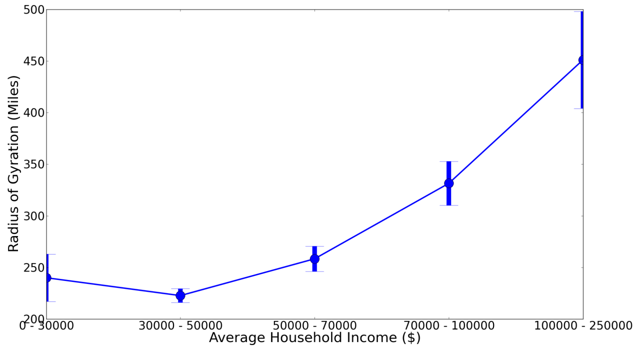

Figure 13: Average R_g versus City Avg Household Income

tent implicitly reveal characteristics of the mobility of users?

6.1 Geographic and Economic Constraints

We begin by illustrating how geographic and economic constraints can influence human mobility patterns as revealed by location sharing services. We focus on users who are located in US cities with a population of more than 4,000. As one type of geographic constraint we consider population density and compare the radius of gyration for users from cities of differing density.[3]

As shown in Figure 12, we can clearly see that people in the densest areas travel much more than people in sparse areas, but that people in the sparsest areas travel farther than people in slightly denser areas. One possible explanation for both of these observations can be that: people living in metropolitan areas have more opportunities to travel for business to distant cities or countries; and people living in sparse areas (small towns) require longer travel to nearby mid-size cities.

Similarly, we can examine the economic properties of a city to understand whether economic capacity inhibits or encourages more travel by its residents. Specifically, we measure the influence of a city's average household income on its residents' radius of gyration, which is plotted in Figure 13. The figure shows that people in wealthy cities travel more frequently to distant places than people in less rich cities. In the meantime, people in cities with the least incomes travel slightly more than people in richer cities.

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

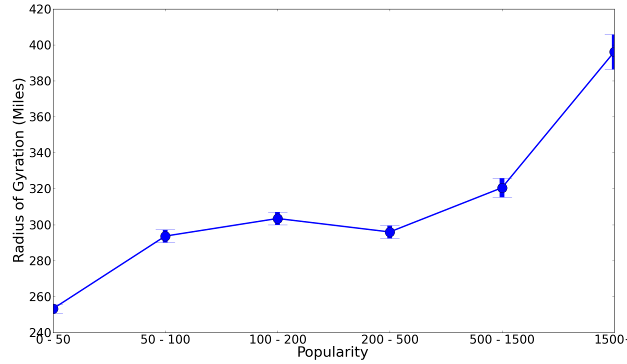

Figure 14: Average R_g versus Popularity

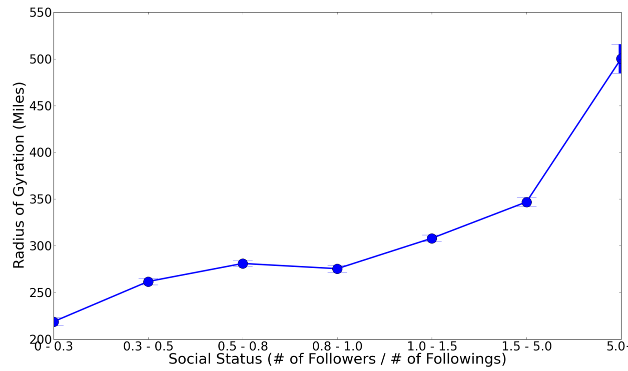

Figure 15: Average R_g versus Social Status

What is encouraging about both these example observations is that location sharing services provide a new window for measuring and studying fundamental properties of cities and their residents.

6.2 Social Status

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

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

High-status users have many followers but follow very few other users themselves. Figure 14 and Figure 15 show the relationship between both of these social status factors and the radius of gyration. We see that in both cases highly social users have higher radii of gyration than less social users. Our initial hypothesis is that users who travel have more chances to meet friends, and thus get involved in more social activities. But perhaps users with lower measured "status" engage with these social media technologies differ-

ently? For example, some Twitter users may primarily only follow other users as a form of news gathering, rather than treating Twitter as a social network of friends, resulting in lower measured status. We are interested to explore these and related questions in our ongoing research.

6.3 Content and Sentiment Factors

Finally, we turn to an analysis of user-generated content in location sharing services and its impact on mobility. Users of location sharing services, in addition to recording their location, can also post short messages, tips, and other annotations on the locations they visit. Unlike purely GPS-driven or cellphone trace data, these short messages provide a potentially rich source of context for better understanding how users engage with location sharing services.

Significant Terms vs. Radius of Gyration: Our first goal is to identify *significant terms* for users associated with varying degrees of radius of gyration, much like in our previous studies of economic, geographic, and social factors. Do high mobility users describe the world differently than low mobility users? We focus our study here on English-language messages only by using the language identification component in the NLTK toolkit (Loper and Bird 2002). We find that 49% of all users (110,559) in our collection are primarily English-language users.

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

In Table 2, we report the top-10 most significant terms from users with different radii of gyration. In the table, we can clearly see the differences between frequent travelers with a large radius of gyration and the more local people with a small radius of gyration. Travelers talk a lot about long journey related terms: "international airport" (and abbreviations of international portals: "SFO", "JFK"), major metropolitan areas (e.g., New York, San Francisco, London, Paris, Los Angeles), "flight", and "hotel". At lower levels of mobility, we see significant words like "railway station" and "bus", as well as discussion of "home", "work", "church", grocery stores (e.g., HEB, Walmart, "mall"), "college", and "university". People with different mobility patterns significantly differ in the topics they talk about and terms they use, indicating a fruitful area of further study.

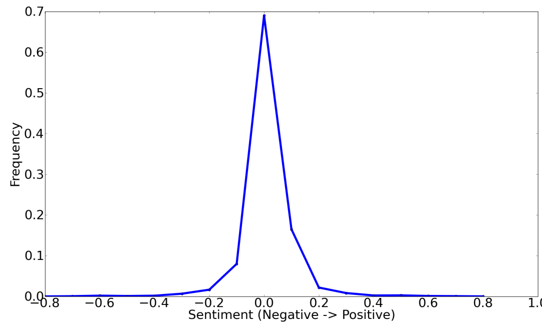

Figure 16: Frequency of Users in Categories of Sentiment

Table 3: Top-10 Significant Terms for Sentiment Category

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

Capturing User's Sentiment: We can additionally measure the relative viewpoint of users and their locations by considering the sentiment of each user's posted messages. To capture the sentiment associated with the checkins, we use the public SentiWordNet (Esuli and Sebastiani 2006) thesaurus to quantify sentiment for each English speaking user. For each message, we extract the words that have a quantified sentiment value in SentiWordNet and consider the sentiment of the post as the mean value for the sentiments for words in the post. For each user, the user's sentiment is calculated as the mean value of the sentiments of all the user's posts. In this way, we capture the sentiment for each of the 110,559 English speaking users in the dataset. The distribution of sentiment of the users is plotted in Figure 16, and we can clearly see that most users have a neutral sentiment, and only a small portion of users express strong sentiment when using location sharing services.

When we drill down to see which words are associated with a positive, neutral, and negative sentiment (again, using mutual information) we see in Table 3 that most of the top neutral terms are likely to be extracted from the auto-generated checkins. In the two categories with non-neutral sentiment, we can clearly see typical words which indicate strong positive and negative sentiment.

However, when we filter the top-100 most positive and most negative terms to only consider location-related terms, we find that there are no location-specific positive terms, but there are many location-specific negative terms. Examples of the words are listed in Table 4. On further inspection of the messages containing these words, we can clearly see the strong negative sentiment associated to the content. For example, when people talk about "MTA", they complain a lot about price increases of MTA's tickets, and its poor service

Table 2: Top 10 Significant Terms for Each Radius of Gyration R_g Category

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

Table 4: Top-20 Location Terms with Negative Sentiment

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

(e.g., “Ticket to the country home has increased by $3. NJ-Transit is worse than the MTA! (@ New York Penn Station w/ 23 others)”, and “I know the MTA is a disaster but 2 of 4 machines being unable to read credit cards at AirTrain station is a new low.”). This preliminary analysis indicates that users are more likely to express negative sentiment about location, and that locations and location-related concepts associated with negative sentiment can be automatically identified based on location sharing services.

7 Conclusion

In this paper, we provide the first large-scale quantitative analysis and modeling of over 22 million checkins of location sharing service users. Concretely, three of our main observations are: (i) LSS users follow simple reproducible patterns; (ii) Social status, in addition to geographic and economic factors, is coupled with mobility; and (iii) Content and sentiment-based analysis of posts can reveal heretofore unobserved context between people and locations. As future work, we are interested to further explore the social structure inherent in location sharing services to study group-based human mobility patterns (e.g., flock behavior). We are also interested in personalized location recommendation based on checkin history and friend-based social mining.

References

Backstrom, L.; Sun, E.; and Marlow, C. 2010. Find me if you can: improving geographical prediction with social and spatial proximity. In *WWW ’10*.

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

Cheng, Z.; Caverlee, J.; and Lee, K. 2010. You are where you tweet: A content-based approach to geo-locating twitter users. In *CIKM ’10*.

Cholera, R.-S. 2011. Words of the year 2010 (the wall street journal). http://on.wsj.com/e7AyTt.

Esuli, A., and Sebastiani, F. 2006. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC’06)*.

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

Foursquare. 2011. So we grew 3400% last year. http://blog.foursquare.com/2011/01/24/2010infographic/.

Golder, S. A.; Wilkinson, D. M.; and Huberman, B. A. 2007. Rhythms of social interaction: Messaging within a massive online network. In *Proceedings of the Third Communities and Technologies Conference*.

Gonzalez, M. C.; Hidalgo, C. A.; and Barabasi, A.-L. 2008. Understanding individual human mobility patterns. *Nature* 453(7196):779–782.

Grove, J. 2010. Foursquare nearing 1 million checkins per day (mashable). http://mashable.com/2010/05/28/foursquare-checkins/.

Hecht, B.; Hong, L.; Suh, B.; and Chi, E. H. 2011. Tweets from justin biebers heart: the dynamics of the location field in user profiles. In *SIGCHI ’11*.

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

Lindqvist, J., et al. 2011. I’m the mayor of my house: Examining why people use foursquare - a social-driven location sharing application. In *SIGCHI ’11*.

Loper, E., and Bird, S. 2002. NLTK: the Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics*.

Rhee, I.; Shin, M.; Hong, S.; Lee, K.; and Chong, S. 2008. On the Levy-Walk nature of human mobility. In *INFOCOM ’08*. IEEE.

Song, C.; Qu, Z.; Blumm, N.; and Barabasi, A.-L. 2010. Limits of Predictability in Human Mobility. *Science* 327(5968):1018–1021.

Yardi, S., and Boyd, D. 2010. Tweeting from the town square: Measuring geographic local networks. In *ICWSM ’10*.

Zheng, Y., et al. 2009. Mining interesting locations and travel sequences from GPS trajectories. In *WWW ’10*.