

Analysis of a Location-based Social Network

Nan Li

Department of Computer Sciences
University of Massachusetts Lowell
1 University Avenue, Lowell, MA 01854
nli@cs.uml.edu

Guanling Chen

Department of Computer Sciences
University of Massachusetts Lowell
1 University Avenue, Lowell, MA 01854
glchen@cs.uml.edu

Abstract—Location-based Social Networks (LSNs) allow users to see where their friends are, to search location-tagged content within their social graph, and to meet others nearby. The recent availability of open mobile platforms, such as Apple iPhones and Google Android phones, makes LSNs much more accessible to mobile users.

To study how users share their location in real world, we collected traces from a commercial LSN service operated by a startup company. In this paper, we present results of data analysis over user profiles, update activities, mobility characteristics, social graphs, and attribute correlations. To the best of our knowledge, this study is the first large-scale quantitative analysis of a real-world commercial LSN service.

I. INTRODUCTION

In the past several years, Online Social Networks (OSNs), such as Facebook and Myspace, have become extremely popular with more than half billion users worldwide. Recently, Location-based Social Networks (LSNs) have also emerged to allow users to see where their friends are, to search location-tagged content within their social graph, and to meet others nearby. The increasing availability of open smartphone platforms, such as Apple iPhones and Google Android phones, makes LSNs much more accessible to mobile users. A recent study shows that the social network sites are among the top mobile Web destinations [21], and there are already many social network applications available on popular Apple iPhones [4]. It is expected that LSN services will attract 82 million subscribers by 2013 [22].

OSN services allow users to easily share thoughts, activities, photos, and other information with friends to strengthen their connections. This kind of sharing of user-generated content, sometimes called status updates or micro-blogging [7], is becoming extremely popular with successful services like Twitter.¹ LSNs take a step further to allow users to share her current location, which can be broadcast to her friends or be used to tag her other shared content. While users may explicitly input their location through a PC-based client, recent smartphone platforms supporting GPS-based or signal triangulation based localization technologies make it much easier for users to access and share their location with LSNs.

So far there have been few studies on how these LSNs are being used by real-world users. It is unclear who are using these services, how they use these services, how they disclose

their locations, how they socialize on the move, and so on. To obtain basic understandings of real-world LSNs, we conducted an extensive study of Brightkite, a commercial LSN service that allows users to share their location, post notes, and upload photos with adjustable privacy settings.

We collected two-month Brightkite data from August to September 2008. Using statistical and data-mining tools, we analyzed the traces including user profiles, activity updates, mobility characteristics, social graphs, and the correlations among different attributes.

Based on the usage patterns and users' location clusters, we found that users' mobility patterns can be classified into four groups: home, home-vacation, home-work, and all other users (§V-B). The social graph for Brightkite is fairly sparse since it is an early-stage service, though the degree distribution still follows power law (§V-C). High-degree users are likely more mobile, have more friends, and send more updates. SMS and Email users are more mobile but their location updates are harder to predict (§V-D1). By clustering over attributes from profiles, activities, mobility, and social graphs, we can classify all users into five groups (§V-D2).

The contribution of this paper is the large-scale analysis of a real-world LSN. In addition to typical studies of social graph structural properties, we studied the correlation between different profile features, activity updates, and mobility patterns to provide deeper insights on user mobility and overall behaviors when using such a LSN. These results can help service providers to understand user behaviors for targeted advertising and construct performance models.

The rest of this paper is organized as follows. In Section II, we discuss the related work. Section III presents background for LSN services, such as Brightkite. Section IV describes how we collected the data traces. All our analysis results are presented in Section V. We give further discussions in Section VI. Finally, we summarize our findings and conclude in Section VII.

II. RELATED WORK

There have been numerous studies on OSNs. Java et al. studied a microblogging service, Twitter, focusing on the topological and geographical properties of its social graph [11]. Mislove et al. studied extensively the structural graph properties of several popular OSN services, such as Flickr, YouTube, LiveJournal, and Orkut [20]. Nazir et al. studied how social

¹<http://www.twitter.com/>

gaming and utility applications were used on Facebook [23]. We believe, however, that this paper differs in that it is the first large-scale quantitative study of a commercial LSN, focusing on both social and mobility properties with graph structures.

The Reality Mining project at MIT studies mobile social systems by distributing to about 100 users mobile phones with customized software that records call history for social links, and Bluetooth scans for geographical proximity [6]. These passively collected measurements were used to compare with users' self-reported social network structures. Our study, however, focuses on mobile social networks of location sharing rather than call records.

Ludford et al. studied how people shared the location knowledge thorough different location types using two small-scale controlled experiments [17]. On the other hand, we studied the users' attributes instead of location types, using a much-larger scale real-world data trace.

One way to study user mobility is to passively observe how users' mobile devices associate with wireless access points (APs), assuming that a device's re-association indicates the user movement. This measurement-based approach has been used to study user mobility in a Computing Science department building [28], on academic campuses [15], [8], [19], on a corporate campus [1], and in a metropolitan area [29]. These studies have shown limited mobility that means users spent most of their time in their home location, while PDA users are more mobile than laptop users [19] and academic users are less mobile than corporate users [8]. To classify user mobility patterns, Kim and Kotz used three features to get six classes: the maximum hourly diameter, and two strongest periods of the DFT transformation of users' vector of hourly diameters [13]. Tang and Baker used nine parameters that resulted in eleven clusters of user mobility patterns [29]. Other researchers have conducted mobility studies using location traces from GPS-equipped vehicles [14], inter-device contact history from specially-designed sensors [3], or class schedules for students [27].

Unlike previous network-measurement based studies, our work uses data collected at the application level where users *voluntarily* share their location. The focus of this paper is not to establish a mobility model from the location trace. Rather, we conduct correlated analysis on mobility from the perspectives of user's social network activities and properties, since we have user profiles and friend lists that are missing in other studies.

III. LOCATION-BASED SOCIAL NETWORKS

Dodgeball² is one of the first LSN service that relies heavily on SMS to allow users to "check in" their current location and to find their friends and friends-of-friends within 10 block radius [10]. On the other hand, Loopt³ leverages GPS and other signal triangulation technologies to automatically sense device location, without requiring manual location updates.

²<http://www.dodgeball.com/>

³<http://www.loopt.com/>

Brightkite⁴ is a Denver-based startup, founded in 2005, that allows users to share their location, to post notes, and to upload photos through a number of interfaces, including Web, SMS, and Email. Recently, the company has also released a native client application on Apple iPhone and is planning a version for Google Android phones. These native client applications, like Loopt clients, leverage GPS and other on-device technologies for automatic location sensing, though still requiring users to hit "check in" button to update location.

Brightkite allows users to define their friends and subscribe to their *activity streams*, including locations they checked in, their posted notes, and their uploaded photos. A note is limited to contain maximum 140 characters, for users to share quick thoughts and short status updates. The "friendship" relation is mutual: a user X accepting Y 's friend request means that X and Y become each other's friend. A user may choose to *protect* her activity stream so only her friends can see her location/note/photo updates. A user may discover nearby people and browse their public activity streams.

The posted notes and the photos are all tagged with user's most recent checked-in location. Once a user checked in at a location, she is assumed to stay at that place until she explicitly checked in at another location. This mechanism gives users a complete control on when and where to share their location, addressing some privacy issues when sharing sensitive location information. When sharing user's current location, Brightkite allows users to control the "granularity." Namely, users can check in at a country, a city, or a zip code, without specifying the exact address.

IV. DATA COLLECTION

Brightkite website pushes *public* (unprotected) activity streams to a data distribution service, GNIP,⁵ from which we collected and archived daily Brightkite data for two months (July 29 to September 29, 2008). The published activity data, however, only contains four fields: the timestamp, user ID, activity type (check in, text note, or photo object), and the Brightkite GUID (Global Unique ID) for that activity. Every day after midnight, our *query* script used the activity GUID, found in previous-day GNIP data, and Brightkite Web APIs to get the location attached to each published activity. Occasionally, a user may post an activity and then remove it shortly after, such as for testing purposes. Thus an activity may appear in GNIP data but our query script could not find that activity on Brightkite website, resulting in a "deleted" entry.

An example XML-represented location information is shown in Figure 1. This place is represented at a city scope without exact street address. The latitude and longitude represent the central point of that city.

Given the IDs of those users who posted at least one update, we used Brightkite Web APIs to retrieve those users' profiles and their list of friends to construct the social graph.

⁴<http://www.brightkite.com/>

⁵<http://www.gnip.org/>

```

<place>
  <scope>city</scope>
  <latitude type="float">39.633141</latitude>
  <longitude type="float">-75.659774</longitude>
  <display_location>Bear, DE, USA</display_location>
  <zip nil="true"></zip>
  <street2 nil="true"></street2>
  <state>DE</state>
  <name>Bear, DE, USA</name>
  <country>US</country>
  <id>19346980e871c96ce0b2d0ecddb6f6666e6786a3</id>
  <city>Bear</city>
  <street nil="true"></street>
</place>

```

Fig. 1. An example of city-scope location representation.

V. TRACE ANALYSIS

Over the two-month period, we collected 300,707 activity updates, including checkin, notes, and photos. Of these updates, 3,996 were removed by the users before our query script could retrieve their attached location. There were 13,512 unique users who sent these non-protected updates. There were 12,228 users who updated their location through checkin, 5,729 users who posted notes, and 2,896 users who posted photos. Table I shows some categorical statistics of the activity updates. About 45% of public activities were checkin updates, when users explicitly updated their current locations. Another 45% of updates were text notes and 10% were photos shared by users.

	categories	updates		users	
types	checkin	135,596	45.1%	12,228	90.5%
	note	134,279	44.7%	5,729	42.4%
	photo	30,832	10.2%	2,896	21.4%
clients	web	227,724	75.7%	12,892	95.4%
	sms	38,536	12.8%	1,849	13.7%
	email	30,232	10.1%	2,425	17.9%
	spot	217	0.1%	14	0.1%
	fireeagle	2	0.0%	1	0.0%
	deleted	3,996	1.3%	1,311	9.7%
scopes	address	126,693	42.1%	6,165	45.6%
	city	100,602	33.5%	8,344	61.8%
	zip	23,952	8.0%	2,592	19.2%
	street	17,986	6.0%	2,266	16.8%
	zip+4	11,374	3.8%	686	5.1%
	state	8,792	2.9%	850	6.3%
	intersection	3,322	1.1%	471	3.5%
	country	3,508	1.2%	505	3.7%
	unknown	482	0.1%	127	0.9%
	deleted	3,996	1.3%	1,311	9.7%

TABLE I
CATEGORIZING 300,707 BRIGHTKITE PUBLIC ACTIVITY UPDATES OVER TWO MONTHS. THERE WERE 13,512 USERS WHO MADE AT LEAST ONE UPDATE.

About 76% updates came from Web, 13% updates came from SMS, and about 10% updates came from Email. There were 95.4% users who used Web, 17.9% users who used Email, and 13.7% users who used SMS for updates. Note that updates from Web/Email may be sent either from PC or

smartphones, and a user may use multiple types of clients to send updates. Apparently, the primary way to access Brightkite is still Web.

A. User profiles

In this section we answer the question of who are using Brightkite by studying the collected user profiles, which contain users' gender, age, and tags. Tags are keywords users put in their profiles to describe, for example, their work and personal interests.

We found that 73% users are male, 17% users are female, and about 10% users did not specify their gender. Clearly Brightkite is dominated by male users, different than the results from other social networks [9], [18].

By analyzing the users' age information, we found that the median age is 27 for males and 25 for females, respectively. Thus female users tend to be younger than male users. There were more than 44% users who were at least 30 years old. On the other hand, only 19% Facebook users were above age 30 [5]. The difference suggests that Brightkite users are more likely to be professionals and explains why men dominate Brightkite, while women dominate most other social networks except LinkedIn for professionals [18]. It is thus not surprising to see that women users are younger than men on Brightkite, which is not the case for other women-dominated social networks [9].

Users may tag their profiles with keywords that describe their jobs and interests. By aggregating these tags, we may have a basic understanding what kinds of people use Brightkite. Table II shows the top 12 tags for different gender and age groups. It is clear that many users are technology oriented as early adopters. Another type of users seem to be interested in marketing over social media. It appears that most users are interested in music. Same is true for books, except for users below 20 who are most likely to be students and more technology oriented. Photography is also a popular tag across all users. It is interesting to see that female users are more into reading, writing, travel, and they are more likely to be bloggers than male users. Social media is a popular tag for both male and female users more than 30 years old, who are likely to be marketing specialists. On the other hand, design and art are more popular among younger users below age 30; and design is more popular for male users while art is more popular for female users.

B. Mobility characteristics

In this section, we study users' mobility characteristics. Unlike previous work on mobility modeling using passive and relatively continuous movement data, the mobility in our study is based on location voluntarily disclosed by users, which is *non-continuous*. Models and patterns derived from such non-continuous data may still be useful to the service provider. For example, the mobility patterns derived from our data is appropriate to build workload model to evaluate location-based information sharing algorithms [2].

	Total Users	Gender		Age			
		male	female	< 20	20 – 30	30 – 40	> 40
1	music	music	music	music	music	music	music
2	books	books	books	apple	books	books	books
3	photography	photography	photography	mac	design	geek	photography
4	design	apple	travel	geek	photography	photography	mac
5	apple	design	art	computers	apple	mac	travel
6	geek	geek	blogger	design	movies	web	technology
7	technology	technology	reading	tech	web	technology	apple
8	web	web	movies	web	geek	apple	blogger
9	mac	mac	writing	student	technology	movies	art
10	movies	movies	design	photography	computers	iphone	iphone
11	social media	social media	writer	books	art	social media	social media
12	travel	computers	social media	art	mac	design	movies

TABLE II
TOP 12 TAGS OF DIFFERENT GENDER AND AGE GROUPS.

We calculated the users' uniquely visited places and the *length* of their movement path. The length is defined as the total number of visited places (may have duplicates) on the path. When calculating the path length, we did not count a newly visited place if it was same as the currently reported place. About 48% users never moved. There were 88% users visited less than 10 unique places and 82% users had path length less than 10, which means that most users did not appear to be very mobile. On the other hand, 1.6% users traveled more than 100 places on their movement path. On the extreme case, a user, who turned out to be a truck driver, visited 413 unique places.

Other than the two mobility metrics focus on number of "hops", we also calculated the users' movement diameter and total distance in miles. Here movement diameter is defined as the maximum distance in miles between any two places visited by the user, and total distance is defined as the summation of the distances of all links on the path. Again, about half of users barely moved while 23% and 10% users had movement diameter more than 100 miles (160 km) and 1000 miles (1600 km), respectively. For total distance, there were 30% and 15% users who had total distance greater than 100 miles (160 km) and 1000 miles (1600 km), respectively. The maximum total distance traveled by a user is 95,155 miles (152,248 km), accumulated by traveling between US, UK, Finland, Austria, Russia, India, and Australia in September, 2008.

We chose 1518 active users who had made at least 50 activity updates, and used a data mining tool (Weka⁶) to cluster each user's geographic positions. Since we did not know a priori the number of location clusters a user may have, we chose an unsupervised clustering algorithm X-Means [24], which is an extended K-means algorithm with efficient estimation of the number of clusters. There are several parameters for X-Means, and we set the minimum cluster number to be 1, maximum cluster number to be 20, maximum number of overall iterations to be 10, and the maximum number of iterations in the K-Means loop to be 1000. The algorithm

outputs each user's geographic position's cluster number and how many positions in each cluster. We found that about 29% users had a single cluster, 16% and 17% users had two and four clusters, while 9% users had three clusters of their visited places. About 28% users had more than four clusters.

We then analyzed the percentage of activity updates users sent from each of their location clusters, from which we roughly classify users into four types. This classification is heuristic, though it can still provide intuitive understandings of users' mobility patterns. Service provider can leverage these patterns to design more effective services corresponding to different user types.

- **Home users:** these users were not mobile and their positions formed a single cluster. We found that 445 users, about 29.31% of all active users, were in this group.
- **Home-vacation users:** these users sent most of their updates from one cluster. In this case, at least 50% updates were from one cluster and no other clusters generated more than 20% updates. We assume that the most active cluster is around user's home location or work place; and the other clusters could be places where users take vacation or make other non-frequent trips. We found that 720 users, about 47.43% of all active users, were in this group.
- **Home-work users:** these users sent most of their updates about equally from two clusters. In this case, about 30% to 60% updates were from top two clusters, while there was no other cluster generating more than 20% updates. We assume that those users used Brightkite at two locations frequently, which likely were near their home and work place. We found that 178 users, about 11.73% of all active users, were in this group.
- **Other users:** these users were those who had different cluster patterns than previous users. For example, some user's positions form more than 5 clusters, each generated similar number of updates. One possibility is that these users only sent updates when they travel, either because they found that home and work areas too familiar to be interesting or because they were concerned about privacy

⁶<http://www.cs.waikato.ac.nz/ml/weka/>

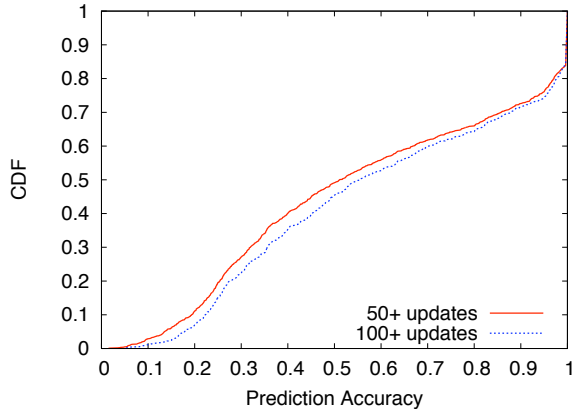


Fig. 2. Distribution of location prediction accuracy.

of disclosing home and work locations. We found that 175 users, about 11.53% of all active users, were in this group.

The next question we want to answer is whether users' movement is predictable. Specifically, given the history of users' movement, how well can we predict the next location where users will send an update? The results of location prediction can be useful to a variety of applications that need, for example, content prefetching or location-based advertisements. We chose active users and ran the Markov-based location predictor developed by Song and Kotz [26]. Their work shows that the $O(2)$ Markov predictor, with fall-back to $O(1)$ whenever it had no prediction, has the best accuracy.

Figure 2 shows the distribution of prediction accuracy for those users who made at least 50 and 100 updates, respectively. The y -axis denotes the cumulative distribution function (CDF) of the prediction accuracy (x -axis). The median accuracy was 49% and there were 17% users whose location could be predicted with at least 99% accuracy. Those users are likely to be those who sent most updates in a single place. For those users who send at least 100 updates, the prediction accuracy improved only slightly.

C. Social graphs

There were 13,512 active users who made at least one activity update in our trace. By aggregating these users' friend lists, we identified 21,941 unique users. Since the Brightkite's friendship is a symmetric relationship, the friend lists allow us to build an undirected social graph. In this paper, we focus our study on those 13,512 active users and ignored their links to other non-active users in their friend lists.

Table III describes the structural properties of Brightkite social graph composed of active users. We also compared these structural properties with the corresponding values for the social graphs of Twitter [11] and a blogosphere WWE (Workshop on Weblogging Ecosystems) [25]. The average degree (number of friends) of Brightkite graph is lower than that of Twitter, possibly because we used an incomplete graph (we removed those non-active users).

Property	Brightkite	Twitter	WWE
Total Nodes	13,512	87,897	143,736
Total Links	55,616	829,247	707,761
Average Degree	8.23	18.86	4.924
(In)degree Slope	-1.9	-2.4	-2.38
Graph Diameter	13	6	12
Clustering Coefficient	0.0782	0.106	0.0632

TABLE III
BRIGHTKITE SOCIAL GRAPH PROPERTIES

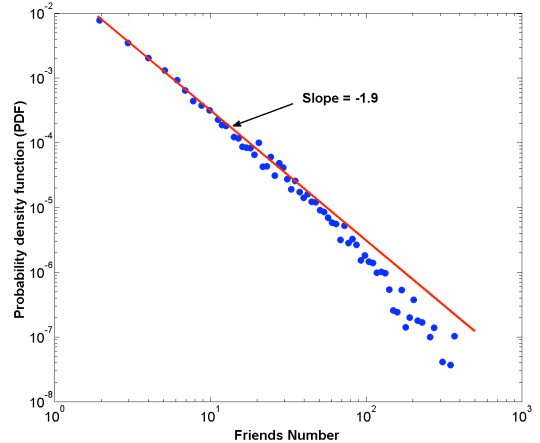


Fig. 3. Degree distribution for every user.

Figure 3 shows the degree distribution of the Brightkite graph, indicating that the distribution follows power law with a fitted-curve having a slope of -1.9 (the solid line). The degree slope for Brightkite graph, however, is smaller than those of Twitter and WWE graphs. Note that the Twitter graph is directional so we only compared with its inbound degree slopes. Twitter graph, however, has the same inbound and outbound degree slopes [11].

Brightkite's graph diameter, defined as the maximum length of the shortest paths between any two nodes, was quite different than that of Twitter probably because of the sparser graph. On the other hand, the clustering coefficient for Brightkite is comparable with that of Twitter. Clustering coefficient is defined as the ratio of the number of links among a node's neighbors divided by the maximum possible number of links among those neighbors. If that ratio is 1, that node's neighborhood is fully connected. If the ratio is closer to 0, that neighborhood is hardly connected.

By analyzing the social graph characteristics, we found that a node with large degree is more likely connected to the nodes that also have large degrees. This suggests that in social networks, the popular users' friends are also tend to be popular. The popular users means that they have more friends than average users. For the users who have more than 10 friends, the average degree (friend number) of their friends is 58.1, while the average degree for all users is 8.23.

To study the correlation of two attributes, we used Spearman's rank correlation coefficient [12] that is defined by

$$\rho \equiv 1 - 6 \frac{\sum d_i^2}{N(N^2 - 1)} \quad (1)$$

Where d_i is the difference between the ranks of corresponding variables, N is the number of variables in each sample set. The coefficient ρ is in the range of $[-1, +1]$. When $\rho = 1$, it means that the two variables have strong positive correlation, and they increase or decrease simultaneously. On the other hand, when $\rho = -1$, it means that the two variables have strong negative correlation, and they increase and decrease in opposite directions. When ρ is near 0, it means there is no obvious correlation between two attributes.

To determine the exact threshold for ρ to suggest significant correlation is related to the size of sample set, which is out of this paper’s scope. In practice, we chose 0.5 and -0.5 as the correlation threshold since we have a large sample set. In many cases, we also do not care the absolute correlation values. Rather, we used the correlation coefficient to compare the correlation strength of different attributes pairs.

Thus we calculated the Spearman’s rank correlation coefficient between a user’s friends number (degree) and the average number of friends for that user’s friends. The coefficient is 0.626, which suggests reasonable positive correlation between the two attributes. This confirms that popular users are more likely connected to other popular users.

In order to uncover the popular users’ activity characteristics, we selected the two groups of popular users who have at least 10 friends and at least 40 friends, respectively, to calculate the average number of total updates. The average number of updates is 53.07 and 77.15, respectively. Comparing with the average number of total updates for all users, which is 22.25, we can conclude that the popular users tend to be more active than average users. In addition, we also calculated the correlation coefficient between the user’s friends number and their number of updates. The coefficient is 0.520, which suggests the same result.

Figure 4 shows the friend number (degree) distribution in different gender groups. Female users had more friends than male users, though a large portion of users had few friends as Brightkite is still an early-stage service.

D. Correlated data analysis

In this section, we study correlation between several user attributes and mobility-related metrics to see what factors impact user mobility. Then we also use multiple attributes to classify active users into different groups, each showing different user behaviors when using the LSN service. Note that this classification is different than the location-clustering based mobility classification (Section V-B, which focused only on mobility patterns.

1) *Correlation coefficient*: We analyzed the correlated relationships between users’ attributes with users’ mobility characteristics. Since not all correlations are linear, we used Spearman’s rank correlation coefficient (Section V-C) to assess how well an arbitrary monotonic function could describe

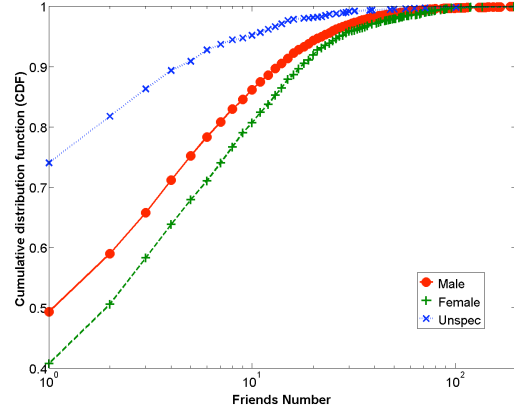


Fig. 4. Degree distributions for male, female, and gender-unspecified users.

the relationship between two variables without making any assumptions about the distributions of those two variables.

Table IV describes the correlation coefficient of multiple attributes with mobility-related metrics. Note that each place on the mobility path corresponds to an activity update, thus we chose $(total_updates - path_length)$ as the first variable and path length as the second variable to calculate the correlated coefficient, which removes the intrinsic correlation between total update with path length.

The number of user’s total updates and active days have clear positive correlation with all mobility metrics except prediction accuracy. This suggests that the more mobile users are, the more active they are and vice versa. The unique location-sharing feature of Brightkite could be one of reasons attracting mobile users. On the other hand, these two attributes have no obvious impact on prediction accuracy.

The next three rows in Table IV describe the correlation of different update clients with mobility metrics. Web proportion is the ratio of a user’s updates that were from Web clients. The Web proportion has slight negative correlation with the first two mobility metrics but almost no relationship with mobility diameter and total distance. On the other hand, it has strong positive correlation with prediction accuracy. This suggests that Web users are less mobile and it is much easier to predict their location. The other two client types (SMS and Email) show opposite characteristics, showing strong positive correlation with mobility metrics except prediction accuracy. This clearly suggests that SMS users are much more mobile. It is interesting to see that Email users are also quite mobile, suggesting that these users likely own smartphones, such as Blackberry and iPhone, and they prefer using Email updates on their phones to using a clunky Web interface.

We also calculated the correlated coefficient among mobility metrics themselves. As expected, the four attributes correlated positively and correlated with prediction accuracy negatively, meaning that it is harder to predict if the user is more mobile.

The last row in Table IV describes the correlation between user’s friends number (degree) with mobility metrics. The

attributes	unique places	path length	mobility diameter	total distance	prediction accuracy
total update	0.808	0.556	0.705	0.732	0.15
active days	0.789	0.805	0.693	0.721	0.04
web proportion	-0.21	-0.215	-0.0096	-0.114	0.608
sms proportion	0.588	0.588	0.559	0.569	-0.232
email proportion	0.619	0.617	0.571	0.586	-0.471
unique places	1	0.992	0.91	0.929	-0.847
path length	0.992	1	0.896	0.921	-0.802
mobility diameter	0.91	0.896	1	0.995	-0.538
total distance	0.929	0.921	0.995	1	-0.617
number of friends	0.515	0.52	0.472	0.489	-0.323

TABLE IV
CORRELATED COEFFICIENT OF MULTIPLE ATTRIBUTES WITH MOBILITY-RELATED METRICS.

result shows positive correlation between friends number and all mobility metrics except prediction accuracy. This suggests that if a user has more friends, she tends to be more mobile, probably visiting her friends at different locations.

Next we discuss correlation of location prediction accuracy with other attributes. There are two main factors impact the prediction accuracy, fewer possible positions and larger sample size. Consider the first factor, it is more difficult to predict a more mobile user's location, which is confirmed by the correlations discussed above. Considering the second factor, we can explain why the total updates and active days have positive correlation with the first four mobility metrics, and the correlation with prediction accuracy is still positive (though small). The large number of updates provides more samples to compensate the accuracy loss due to high mobility.

2) *User classification*: We also used Weka to cluster all users based on multiple user attributes and then categorized these users into different groups given different clustering characteristics.

We chose expectation-maximization (EM) as the clustering algorithm, which is a well-known unsupervised clustering algorithm that is used in statistics for finding maximum likelihood estimates of parameters in probabilistic models. We ran the EM algorithm on eight attributes: ① Total updates, ② Activity days, ③ Retained user (this attribute is a binary value that indicates whether the user updated Brightkite in two continuous weeks) [11], ④ Uniquely visited places, ⑤ Path length, ⑥ Mobility diameter (miles), ⑦ Total distance (miles), ⑧ Number of friends.

The EM algorithm returned 5 clusters and Table V shows the results of each cluster.

The cluster 1 consists of 30% users, we consider that they were *inactive users*, since there were on average only 7.19 updates in 64 days and the values of other attributes were all less than average.

The cluster 2 consists of 16% users. We believe that they were *normal users*. They login the Brightkite every two days on average, 99.3% are "loyal" users, and their mobility diameter and total distance are closer to the average values than users in other clusters.

The cluster 3 consists of 6% users. They were *active users* because their average number of updates was much larger than those of users in other clusters. Same was true for all the other

attributes except mobility diameter and total distance.

The cluster 4 consists of 8% users. They were *mobile users*, because their mobility diameter and total distance were distinctly large. We assume that those users traveled frequently and kept using Brightkite to share their location.

The cluster 5 consists of 41% users, which were *trial users*. They updated only once or twice and never came back. There were no loyal users in this group.

VI. DISCUSSIONS

We found that one user traveled 95,155 miles in 62 days, which means that he traveled 1500 miles per day on average. The location checkins did show that he visited several countries in a single month. We could not, however, verify that he actually visited these places by reading his blog website, checking his microblogs on Twitter, and browsing his photos on Flickr. Since Brightkite does not verify user's checkin locations, the user-disclosed data may be intentionally forged. On the other side, location disclosure is more sensitive than traditional sharing of information. How to share the location information while keeping users' privacy is a challenging research question.

Due to Brightkite's data publishing policy, our traces do not include the *protected* user updates, though previous studies have shown that few social network users change their default privacy settings (usually public) [16]. Unfortunately we do not know how many Brightkite users actually protect their updates, and the results in this paper can only be reliably applied to users with public updates. In Section V-D2, we found about 41% users are trial users. It is possible that some of them tried Brightkite at first and changed into protected mode for daily using. This hypothesis, however, remains to be confirmed.

VII. CONCLUSION AND FUTURE WORK

In this paper, we present results of a large-scale quantitative analysis of Brightkite, a commercial location-based social network (LSN). Unlike other social networks, Brightkite is dominated by male users who are professionals and likely to be bloggers and work in social media area. On the other hand, women users are younger than their male peers.

Based on the patterns of users' location clusters, we can classify users' mobility patterns into four mobility groups. The

Cluster	User Type	Proportion %	Total Updates	Active Days	Retained %	Unique Places	Path Length	Mobility Diameter	Total Distance	Friends Number
	average	N/A	22.25	8.52	56.94	5.41	9.58	388.49	835.05	8.23
1	inactive	30	7.19	4.56	90.29	2.87	3.57	119.46	173.63	6.47
2	normal	16	47.10	22.07	99.30	10.00	19.02	226.33	624.00	15.69
3	active	6	197.73	47.25	99.60	34.97	77.95	1126.22	3747.30	36.26
4	mobile	8	20.43	10.37	90.90	9.06	12.22	3868.47	7304.41	11.77
5	trial	41	1.23	1	0	1.06	1.06	0.40	0.04	2.35

TABLE V
EM CLUSTERING RESULTS

social graph for Brightkite is fairly sparse since it is an early-stage service, though the degree distribution still follows the power law. High-degree users are likely more mobile, have more friends, and send more updates. SMS and Email users are more mobile and their location updates are harder to predict. By clustering the attributes from profiles, activities, mobility, and social graphs, we can classify all users into five distinct behavior groups.

In future work, we plan to expand our studies on social graphs using additional metrics and combine them for correlated analysis. We also plan to extract a workload model from the location updates to evaluate the performance of location-based information-sharing and privacy-preserving algorithms.

REFERENCES

- [1] Magdalena Balazinska and Paul Castro. Characterizing Mobility and Network Usage in a Corporate Wireless Local-Area Network. In *Proceedings of the First International Conference on Mobile Systems, Applications, and Services*, pages 303–316, San Francisco, CA, May 2003.
- [2] Ying Cai and Toby Xu. Design, analysis, and implementation of a large-scale real-time location-based information sharing system. In *Proceedings of the Sixth International Conference on Mobile Systems, Applications, and Services*, Breckenridge, CO, June 2008.
- [3] Augustin Chaintreau, Pan Hui, Jon Crowcroft, Christophe Diot, Richard Gass, and James Scott. Pocket switched networks: Real-world mobility and its consequences for opportunistic forwarding. Technical Report UCAM-CL-TR-617, University of Cambridge Computer Laboratory, February 2005.
- [4] Guanling Chen and Faruq Rahman. Analyzing privacy designs of mobile social networking applications. In *Proceedings of the IEEE/IFIP International Symposium on Trust, Security and Privacy for Pervasive Applications (TSP)*, pages 83–88, Shanghai, China, December 2008.
- [5] Matt Dickman. The face of Facebook: A marketer's guide to understanding the population of Facebook, October 2008.
- [6] Nathan Eagle, Alex (Sandy) Pentland, and David Lazer. Inferring social network structure using mobile phone data. In *Social Computing, Behavioral Modeling, and Prediction*. Springer, March 2008.
- [7] Shraavan Gaonkar, Jack Li, Romit R. Choudhury, Landon Cox, and Al Schmidt. Micro-blog: sharing and querying content through mobile phones and social participation. In *Proceedings of the Sixth International Conference on Mobile Systems, Applications, and Services*, pages 174–186, Breckenridge, CO, June 2008.
- [8] Tristan Henderson, David Kotz, and Ilya Abyzov. The changing usage of a mature campus-wide wireless network. In *Proceedings of the Tenth Annual International Conference on Mobile Computing and Networking*, pages 187–201, Philadelphia, PA, September 2004.
- [9] Auren Hoffman. Men are from video games, women are from social networks. Rappleaf Blog, March 2008.
- [10] Lee Humphreys. Mobile social networks and social practice: A case study of dodgeball. *Journal of Computer-Mediated Communication*, 13(1):341–360, October 2007.
- [11] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we Twitter: Understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, pages 56–65, August 2007.
- [12] Maurice Kendall and Jean D. Gibbons. *Rank Correlation Methods*. A Charles Griffin Title, 5 edition, September 1990.
- [13] Minkyong Kim and David Kotz. Periodic properties of user mobility and access-point popularity. *Journal of Personal and Ubiquitous Computing*, 11(6):465–479, August 2007.
- [14] Takehiko Kobayashi, Noriteru Shinagawa, and Yoneo Watanabe. Vehicle mobility characterization based on measurement and its application to cellular communication systems. *IEICE Transactions on Communications Special Issue on Multimedia Mobile Communication Systems*, E82-B(12), December 1999.
- [15] David Kotz and Kobby Essien. Analysis of a campus-wide wireless network. In *Proceedings of the Eighth Annual International Conference on Mobile Computing and Networking*, pages 107–118, Atlanta, GA, September 2002.
- [16] Balachander Krishnamurthy and Craig E. Wills. Characterizing privacy in online social networks. In *Proceedings of the Workshop on Online Social Networks*, Seattle, WA, August 2008.
- [17] Pamela J. Ludford, Reid Priedhorsky, Ken Reily, and Loren Terveen. Capturing, sharing, and using local place information. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1235–1244, New York, NY, USA, 2007. ACM.
- [18] Richard MacManus. Study: Women outnumber men on most social networks. ReadWriteWeb.com, July 2008.
- [19] M. McNett and G. M. Voelker. Access and mobility of wireless PDA users. *Mobile Computing Communications Review*, 9(2):40–55, April 2005.
- [20] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the Seventh ACM Internet Measurement Conference*, San Diego, CA, October 2007.
- [21] Social networking and craigslist top mobile destinations, says openwave. Reporter Wireless and Mobile News, March 2009.
- [22] 82 million location-based mobile social networking subscriptions by 2013. ABI Research Market Analysis Report, November 2008.
- [23] Atif Nazir, Saqib Raza, and Chen-Nee Chuah. Unveiling Facebook: A measurement study of social network based applications. In *Proceedings of the Seventh ACM Internet Measurement Conference*, Vouliagmeni, Greece, October 2008.
- [24] Dan Pelleg and Andrew Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 727–734, San Francisco, 2000.
- [25] X. Shi, B. Tseng, and L. A. Adamic. Looking at the blogosphere topology through different lenses. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007)*, March 2007.
- [26] Libo Song, David Kotz, Ravi Jain, and Xiaoning He. Evaluating location predictors with extensive Wi-Fi mobility data. In *Proceedings of the 23rd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, volume 2, pages 1414–1424, March 2004.
- [27] Vikram Srinivasan, Mehul Motani, and Wei Tsang Ooi. Analysis and implications of student contact patterns derived from campus schedules. In *Proceedings of the 12th Annual International Conference on Mobile Computing and Networking*, pages 86–97, Los Angeles, CA, September 2006.
- [28] Diane Tang and Mary Baker. Analysis of a local-area wireless network. In *Proceedings of the Sixth Annual International Conference on Mobile Computing and Networking*, pages 1–10, Boston, MA, August 2000.
- [29] Diane Tang and Mary Baker. Analysis of a metropolitan-area wireless network. *Wireless Networks*, 8(2/3):107–120, March–May 2002.