We Know Where You Live: Privacy Characterization of Foursquare Behavior

Tatiana Pontes*, Marisa Vasconcelos*, Jussara Almeida*, Ponnurangam Kumaraguru†, Virgilio Almeida*

*Universidade Federal de Minas Gerais, Brazil
†Indraprastha Institute of Information Technology, India
*{tpontes,marisav,jussara,virgilio}@dcc.ufmg.br
†pk@iiitd.ac.in

ABSTRACT

In the last few years, the increasing interest in location-based services (LBS) has favored the introduction of geo-referenced information in various Web 2.0 applications, as well as the rise of location-based social networks (LBSN). Foursquare, one of the most popular LBSNs, gives incentives to users who visit (check in) specific places (venues) by means of, for instance, mayorships to frequent visitors. Moreover, users may leave tips at specific venues as well as mark previous tips as done in sign of agreement. Unlike check ins, which are shared only with friends, the lists of mayorships, tips and dones of a user are publicly available to everyone, thus raising concerns about disclosure of the user's movement patterns and interests. We analyze how users explore these publicly available features, and their potential as sources of information leakage. Specifically, we characterize the use of mayorships, tips and dones in Foursquare based on a dataset with around 13 million users. We also analyze whether it is possible to easily infer the home city (state and country) of a user from these publicly available information. Our results indicate that one can easily infer the home city of around 78% of the analyzed users within 50 kilometers.

Author Keywords

Location Prediction, Privacy, Foursquare

ACM Classification Keywords

K.4.1 Computing Milieux: Computers and society—*Public policy issues*

General Terms

Experimentation, Measurement

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
UbiComp '12, Sep 5-Sep 8, 2012, Pittsburgh, USA.
Copyright 2012 ACM 978-1-4503-1224-0/12/09...$10.00.

INTRODUCTION

Online social networks (OSN), such as Facebook, Twitter and the recent Google+ , are currently very popular. Some reasons for their great popularity include the easiness at which users can communicate and share content at large scale, the opportunity for self-promotion, commercial interests, as well as the simple intent of socialization [16]. Thus, users share a lot of information about themselves including age, address, relationship status, photos, and topics of interests on OSNs.

Due to the increasing use of smart devices equipped with Global Positioning System (GPS), LBSs have become very prevalent, thus attracting the interest of the research community. They have also motivated the creation of LBSNs [20], which emerge with an additional attraction in relation to OSNs, namely, the association of geographical information with the shared data. Out of the various existing LBSNs, such as Gowalla and Brightkite, Foursquare[1] is currently one of the most popular ones. Its overall goal revolves around the location sharing while users accumulate awards for visiting specific places in the system. It has been recently reported that Foursquare already achieved over 20 million members, with a history of around two billion visits notified by users in places all over the world.[2]

In Foursquare, users can inform their friends about their current location through *check ins* which may be converted into virtual rewards as badges or mayorships if the user is a frequent visitor of the same venue. Besides this gamification aspect, Foursquare has massively invested into the recommendation aspect allowing users to leave notes (tips) for friends and other users about their experiences at specific places (venues). Users can also keep track of tips marking them as done or saving them in a to-do list.

Easy availability of information about the location of a user raises several concerns about privacy violation [18]. For instance, the information about one's location may facilitate inferences about her behavioral patterns and habits. For instance, in Foursquare, although check ins are shared only with the user's friends, the use of other features of the system, such as mayorships, tips and dones are publicly avail-

[1] http://foursquare.com
[2] http://mashable.com/2012/04/16/foursquare-20-million/

able to everyone. In other words, the information about the venue(s) where the user is mayor (if any) as well as all venues where she left a tip or marked a tip as done or to-do is available to anyone. This information may reveal a lot about a user. For instance, a mayorship at a specific venue means that the user is a frequent visitor of that venue, whereas a tip (or a done/to-do) implies a prior visit or intention to checking out the place in the future. Tipping patterns may ultimately reveal user habits and personal interests. Indeed, if one considers that writing a tip requires more effort from the user than simply doing a check in, it could be argued that the locations at which a user left tips are even stronger indications of places she actually visited than check ins.[3] Users may do check ins when they are traveling, far from home, to show their friends that they are enjoying different places.

In this paper, we analyze how users explore these publicly available features, notably, mayorships, tips and dones in Foursquare, and their potential as source of information leakage and privacy violation. More specifically, we provide a characterization of mayorships, tips and dones in Foursquare based on a large dataset we crawled containing information on more than 13 million users and 15 million venues. As a first step towards investigating how much information about a user can be inferred from her tips, dones and mayorships, we analyze whether one can easily infer the home city (state and country) of a user from these publicly available information, by simply taking the location of the majority of the venues the user is connected to via mayorships, tips and dones. Note that the home city in Foursquare user profile is not a mandatory field and appear as an open text field. Thus, a user may choose not to reveal her home city by simply writing an invalid city name or even leaving it blank. Recent analyses of the location field in Twitter have pointed out that 34% of the users did not provide real locations, often including fake locations or even sarcastic comments. One of the reasons that justifies this user behavior may be to avoid unwanted messages that, for instance, may use the location information to provide a more efficient targeted advertisement mechanism. The question that we address here is: *despite being a private data that the user may choose not to reveal, can we still infer the home city of a user in Foursquare from her mayorships, tips and dones?*

We note that the literature contains several models for predicting user's home city mainly, exploiting the contents of user messages [1, 8, 11] or location of their friends [6]. Focused mainly on Twitter, these prior efforts aim at improving personalized services [1], performing targeted regional advertisements [19] or even detecting major events [15]. Instead, we here focus on a different application, Foursquare, exploiting different publicly accessible features, as our intention is to investigate their potential as source of inference of information about the user.

RELATED WORK

The increasing popularity of LBSNs have attracted researchers towards the awareness of location data. A number of recent studies have focused on geographically referenced information addressing aspects such as understanding why users share their location [16], human mobility patterns [2, 3, 14], user profile identification [10, 17], event detection [15] and analysis of a city urban development through check ins [5].

The information sharing in LBSNs and online social networks in general also raises concerns about exposure of user private data, touching privacy related issues. For instance, some studies have shown that it is possible to infer user implicit data through explicit information shared in such systems [7]. Mislove *et al.* have shown that users' personal interest can be inferred from friends [12], specially because, as argued in [4], people with common preferences are more likely to be friends. Other studies focused on assessing how users face privacy related issues and which strategies they often adopt to manage their exposure in the system [9].

There have also been studies that investigate whether it is possible to infer a user's location through other features which contain implicit location information. In [6], the home location of Twitter users are inferred from friends, with the simple assumption that users tend to have friends that live near them. In [1], Cheng *et al* estimated the user home city using the content of tweets with the assumption that people who live nearby do have a similar vocabulary. Other studies use machine learning approaches to infer user home location exploiting tweets' textual content [8] or users' tweeting behavior [11]. Unlike these previous studies, we here focus on inferring user's home city in a very popular LBSN (Foursquare), exploiting publicly available features such as mayorships, tips and dones that are associated with location information. To our knowledge, no previous work has addressed this problem yet.

FOURSQUARE DATASET

In this section, we briefly review the main elements of Foursquare as well as the crawled dataset used in our experimental evaluation.

Foursquare: Background

Foursquare is currently the largest and the most popular LBSN where members can share their locations with friends and followers through *check ins*. Check ins are performed via devices with GPS when a user is close to specific locations (*venues*). Venues are pages in the system that represent real locations of a great variety of categories such as airports, hotels, restaurants, monuments or squares.

Foursquare has a playful aspect that gives incentives to users who share more locations. Thus, check ins can be accumulated and exchanged by *badges* and *mayorships*. Badges are like medals given to users who check in at specific venues or achieve some predefined number of check ins. A mayorship, in turn, is given to the user who was the most frequent visitor (in number of check ins) of a venue in the last 60 days. Venue mayors are often granted rewards, promotions, discounts or even courtesies by business and marketing managers who own the venue. Once a user becomes a mayor of a given venue, that mayorship will be listed in her history,

[3]Note that Foursquare allows a user to check in at a venue even if she is not near the corresponding physical location.

even if some other user later ousts her from that position. That is, each user maintains a history of all mayorships she conquered. Multiple mayorships at the same venue are listed only once in this history. Moreover, mayorships are not temporal referenced.

Users can post *tips* at specific venues, commenting on their previous experiences when visiting the corresponding physical places. Tips can also serve as feedback, recommendation or review to help other users choose places to visit. Examples of tips include the best option of a menu in a restaurant, the best place to have lunch in an airport, or even a complaint about a service. With a limitation of 200 characters, tips nourish the relationship between users and real businesses and may be a key feature to attract future visitors [17]. Each user has a history of all tips she posted, with associated venue and timestamp. When visiting a venues' page, after reading a previously posted tip, a user may mark it as *done* or *to-do*, in sign of agreement with the tip's content or intention to visit that location in the future, respectively. The history of mayorships as well as the list of tips and dones, along with corresponding venue and timestamp information, of a user are publicly available at the user's profile page.

Crawled Dataset

Our study is based on a large dataset collected from Foursquare using the system API. We crawled user profile data consisting of user type, user home city, list of friends, mayorships, tips, dones, total number of check ins, Twitter screen names and Facebook identifiers. Our crawler ran from August to October 2011, collecting a total of 13,570,060 users, which is a good approximation of the entire Foursquare community at the time of the crawling since, reportedly, the number of users registered in Foursquare was 10 million in June 2011, reaching 15 million in December of the same year [13]. Our dataset contains 10,618,411 tips, 9,989,325 dones and 15,149,981 mayorships at 15,898,484 different venues.

FOURSQUARE CHARACTERIZATION

In this section, we discuss characterization of Foursquare users, focusing on user attributes that are publicly available in the system API and are associated with geo-referenced information, i.e., home cities, mayorships, tips and dones. Recall that the user home city and the venue location are open text fields, whose validity is not enforced by the system. Indeed, they may carry noise and invalid locations. Thus, we start our study by analyzing the amount of valid location information in our dataset. Next, we analyze the use of tips, dones and mayorships, focusing on the distribution of associated locations around the globe. Finally, we perform a temporal and spatial analysis of user activities in terms of tipping and marking previous tips as done.

Location Information in Foursquare

We here discuss the location information available in public attributes of Foursquare users, i.e., in home city, mayorships, tips and dones.[4] Since mayorships, tips and dones are associated with venues, we here analyze the user home city and the venue location attributes in our dataset.

User home city, in particular, is limited to 100 characters and is not required to be filled. It is expected that users provide the name of the city where they live, although the system provides neither rule to enforce it nor any automatic tool to help users filling the field (e.g., a predefined list of cities from which the user can choose one). Thus, users are free to provide this location information at various granularities, ranging from specific addresses, to city, state and country names, or even regions of the planet (e.g., "North Pole"). We also observed some home city fields filled with emails, phrases, or even numbers in our dataset. Similarly, the location associated with a venue, and thus, indirectly, with mayorships, tips and dones of that particular venue, is also an open text field. Unlike the user home city, the address and the city of a venue must be filled before the venue is created. Moreover, it is necessary to set a pin in a map to update the venue's location. Once again users may choose to provide invalid addresses and city names, and mark arbitrary locations in the map.

Table 1. Dictionary. GI = geographic information. UHC = User Home City. VL = Venue Location.

Statistics	UHC	VL
# in dataset	13,570,060	15,898,484
# valid GI	13,299,112	11,683,813
# valid but ambiguous GI	359,543	2,868,636
# non-GI	244,233	4,214,671
# empty entries	26,715	0

Table 2. Quality of Geographic Information.

Quality	# Users	# Venues
Continent	107	61
Country	602,932	294,596
State	390,224	93,513
County	251,383	276,097
City	10,354,058	6,937,523
Neighborhood	981,139	1,060,124
Area of Interest/Airport	27,307	47,896
Street	326,751	95,543
Point of Interest	5,607	9,792
Coordinate	61	32

Thus, in order to standardize the home city and venue location fields, we created a dictionary of city names using the *Yahoo! PlaceFinder*, the Yahoo's geo-coding API.[5] This tool was used to verify the validity of the data in both fields. For a given query (text), the tool either returns some geographic data, in case the query consists of a valid location, or an error, otherwise. For queries consisting of valid locations, the tool's response depends on the "quality" of the query, which, in turn, is related to the spatial granularity (e.g., street, city, state, country) of the location information provided in the query. For instance, for a query "New York", *Yahoo! PlaceFinder* returns that the query's quality is at the granularity of city, and provides the corresponding geographic coordinates, a standardized city name as well as the state and country names. *Yahoo! PlaceFinder* may also identify locations at the finer granularity of streets. Moreover, note that the use of standardized city name allows us to

[4]Check ins are private, it is not possible to access the geographic location associated with them.

[5]http://developer.yahoo.com/geo/placefinder/

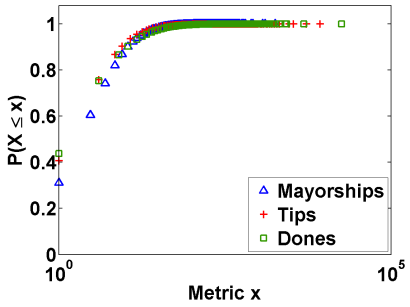


Figure 1. Cumulative Distribution of the Number of Mayorships, Tips and Dones per User.

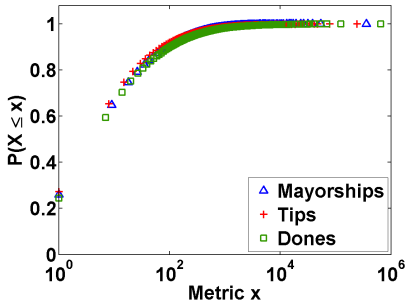


Figure 2. Cumulative Distribution of the Number of Mayorships, Tips and Dones per City.

uniquely identify the city, despite the existence of multiple name variations (e.g., NY, New York City, etc).

Table 1 provides some details about our dictionary, indicating the total number of users and venues with valid, invalid as well as empty location information. Note that, perhaps surprisingly, the vast majority (98%) of the users do provide valid locations, according to *Yahoo! PlaceFinder*, in their home city attributes, and only a tiny fraction of users leave this attribute empty (0.2%). The fraction of venues with valid locations is smaller (73.5%), but, also accounts for most venues in our dataset. We note that, for some queries, *Yahoo! PlaceFinder* returned multiple ambiguous answers reflecting alternative locations with the same name (e.g., there are ten cities named "Springfield" in the United States). We chose to disregard users and venues with ambiguous locations, which correspond to 2.7% and 24.6% of all users and venues with valid locations, respectively, in our dataset.

Next, we analyze the "quality", in terms of spatial granularity, of valid (unambiguous) locations associated with users and venues. In Table 2, we present the distributions of users and venues across 10 different quality levels, ranging from continent to specific coordinates. Note that, the majority of users and venues provide location information at the granularity of city or at finer granularities. Indeed, users and venues are associated with 100,629 different cities around the world. Note, however, that over 1.2 million users provide location information at a coarser granularity, often at the country level. Thus, the inference of the home city or even state of these users based on their mayorships, tips and dones will reveal private information.

Mayorships, Tips and Dones

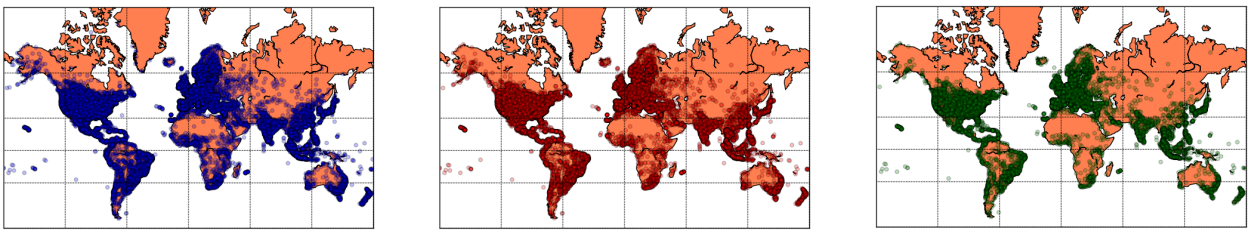
In this section, we analyze the mayorships, tips and dones of users in our dataset. Since our goal is to exploit the location of the venues associated with these attributes to infer the user home city, we start by showing an overview of the use of mayorships, tips and dones among users in our dataset. We observe that almost 4,2 million users, or around 30% of all users in our dataset, have at least one of these attributes. Out of these, around 1 million have only mayorships, 670 thousand have only tips and 367 thousand have only dones, whereas 890 thousand users have all three attributes. Thus, exploiting these attributes to infer a user home city is promising as the required information is available in a large fraction of all users. Moreover, as shown in Figure 1 and consistent with previous analyses of Foursquare [14, 17], the distributions of the numbers of mayorships, tips and dones per user are very skewed, with a heavy tail, implying that few users have many mayorships (tips or dones) while the vast majority have only one mayorship (tip or done). Indeed, for users that have one of these attributes, we find that 69% (59% and 56%) of the users have 2 or more mayorships (tips and dones).

Figure 2 shows the distributions of numbers of mayorships, tips and dones per city, considering only cities with at least one instance of the attribute. As shown, the distributions are also very skewed, with a few cities having as many as 100 mayorships, tips or dones.

Next, we analyzed the correlation between the number of mayorships, tips and dones per city. We found that there is a high correlation between the number of mayorships and the number of tips across cities, with a Spearman's correlation coefficient ρ [21] equal to 0.78. Similarly, the correlation is also high between the number of mayorships and the number of dones ($\rho = 0.72$). Moreover, we found that the cities with the largest numbers of mayorships tend also to have large numbers of tips and dones, although some interesting differences are worth noting. For instance, mayorships are more concentrated in Southeast Asia, in cities like Jakarta, Bandung and Singapore, which are the top three cities in number of mayorships, jointly having more than 500,000 mayorships. Tips, in turn, are concentrated in different locations around the Earth: the top three cities in number of tips are New York, Jakarta and São Paulo, with a total of 600,000 tips. Dones, on the other hand, tend to be concentrated in venues in the United States, in cities like New York, Chicago and San Francisco, which jointly received around 1 million dones.

We note that, although other studies [1, 8, 11] have exploited textual features to analyze user location, we here chose not to exploit the tip's content as they are often targeted towards more generic topics such as food and service quality. We observe that most words extracted from tips in our dataset are adjectives or are related to food, meal, services and generic places where one can eat or drink.

We now discuss the distribution of cities with venues where users have mayorships, tips and dones around the world.



(a) Mayorships. (b) Tips. (c) Dones.

Figure 3. Global Distribution of Mayorships, Tips and Dones.

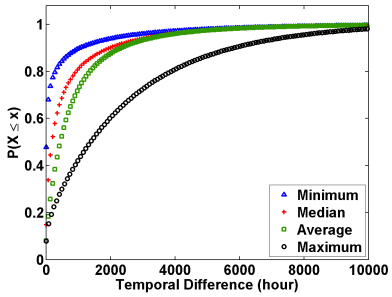


Figure 4. Cumulative Distribution of Time Interval Between Consecutive Tips/Dones Posted per User.

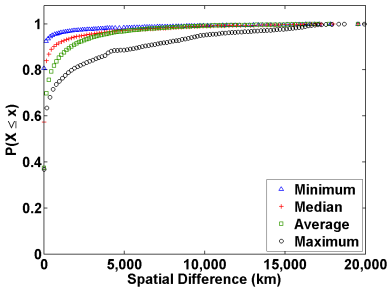


Figure 5. Cumulative Distribution of Displacements Between Consecutive Tips/Dones Posted per User.

We only consider attributes associated with venues that have valid cities (validated by *Yahoo! PlaceFinder*) as location. Figure 3 shows these distributions in maps of the globe, with each point representing a city with venues with at least one mayorship, tip or done.[6] As the maps show, Foursquare venues are spread all over the world, including remote places such as Svalbard, an archipelago in the Arctic Ocean, with coordinates (78.218590,15.648750). Moreover, all three maps are very similar, with most incidences of points in America, Europe and Southeast Asia. The distribution of mayorships, shown in Figure 3(a), is denser, with a total number of unique cities (79,194) much larger than in the distributions of tips and dones, which cover a total of 54,178 and 30,530 unique cities, respectively. The somewhat sparser tip map (Figure 3(b)) indicates that there are many cities, particularly in Canada, Australia, central Asia and Africa, where, despite the existence of venues and mayors, users do not post tips. The distribution of dones, shown in Figure 3(c), reveals an even sparser map, with most activity concentrated in touristic or developed areas, such as USA, western Europe and southeast Asia. We note that a similar map was produced for check ins in [2]. Besides both datasets were collected at different times, we can see that their main areas of concentration overlap.

Temporal and Spatial Analyses

We perform a temporal and spatial analysis of user activity in terms of tips and dones. Our goal is to analyze how often users leave tips / dones as well as how far users "go" between consecutive tips / dones. To that end, we make use of the timestamp associated with each tip and done as well as the location of the venue where the tip (or done) was left.

[6]The Antarctica continent was omitted because there was no point on it.

We start by investigating the frequency at which users leave tips and/or mark previous tips as done. We do so by analyzing the time interval between consecutive activities (be it a tip or a done) of the same user. Thus, we consider only users with at least two activities, covering a total of 1,959,647 users. We summarize user activity by the minimum, median, average and maximum inter-activity times. Figure 4 shows the cumulative distributions of these four measures computed for all considered users. We note that the distribution of minimum inter-activity times is very skewed towards short periods of time, with almost 50% of the users posting consecutive tips/dones 1 hour apart. However, on average, median and maximum, users do tend to experience very long periods of time between consecutive tips and dones. For instance, around 50% of the users have an average inter-activity time of at least 450 hours, whereas around 80% of the users have a maximum inter-activity time above 167 hours (roughly a week).

Next, we analyze the displacement between two venues visited in sequence by the user, as indicated by consecutive tips and/or dones of the user. For this analysis, we consider only users with at least two activities, provided that the venues associated with these activities have valid locations, with "quality" of city level or finer granularity. Our dataset contains almost 1.5 million users in this group. For these users, we computed the displacements between consecutive tips/dones by taking the difference between the coordinates of the associated venues. Once again, we summarize user activity computing the minimum, median, average and maximum displacement per user. Figure 5 shows the distributions of these measures for all analyzed users. Around 36% of the users have average and maximum displacements of 0 kilometer, indicating very short distances (within a few

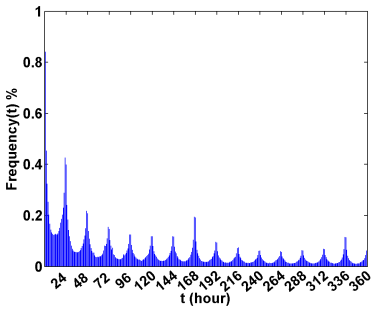


Figure 6. Distribution of Returning Times.

meters). Moreover, 70% of the users have an average displacement of at most 150 kilometers, which could be characterized as within the metropolitan area of a large city. Also 60% of the users have a maximum displacement of at most 100 kilometers, possibly the distance between neighboring cities. Thus, overall, consecutive tips/dones of a user are often posted at places near each other. However, there are exceptions. About 10% of the users have a maximum displacement of at least 6,000 kilometers.[7]

Finally, we analyze how often users return to the same venue for tipping or marking tips as done. That is, we analyze the returning times, defined as the time interval between consecutive tips/dones posted at the same venue by the same user. This analysis is focused on 813,607 users, who have at least two tips/dones in the same venue, and cover more than 3 million returns. We here choose to show the distribution of all measured returning times, as opposed to summarizing them per user first, so as to compare our results against previous findings of check in patterns [2]. Figure 6 shows the distribution, focusing on returning times under 360 hours, which account for 69.7% of all measured observations. The curve shows clear daily patterns with returning times often being multiples of 24 hours, which is very similar to the distribution of returning times computed based on check ins [2]. We note, however, that 50% of the measured returning times are within 1 hour, which cannot be seen in the Figure as its y-axis is truncated at 1% so that the rest of the curve could be distinguished. Moreover, out of these observations, 90% of them are at most 10 minutes. Thus, returning times, in general, tend to be very short. If we analyze the behavior per user (omitted more details, due to space constraints), we note that most users have very short minimum returning times, which is below 1 hour for 62% of the users. However, consistently with results in Figure 4, on average, median and maximum, users do tend to experience longer returning times. For instance, 52% of the users have average returning times of at least 168 hours.

INFERRING USER'S HOME LOCATION

In this section we investigate whether one can infer, with reasonable effectiveness, the location where a user lives based only on information that is publicly available on her Foursquare profile page, notably the lists of mayorships, tips and dones.

[7]Note that the maximum displacement between two points in the Earth is the distance between antipodes (two diametrically opposed points) that is about 20,000 kilometers.

We here discuss the inference approach and evaluation methodology adopted in Methodology section whereas our main results are discussed in Experimental Results section.

Methodology

The key assumption behind this work is that users tend to have mayorships, tips and dones in venues at the same location (e.g., city) where they live. At first, one might think that the mayorship locations are perhaps the strongest piece of evidence about a user's home location, as the former represent places the user possibly goes very often. Recall that a user only becomes mayor of a venue if she is the most frequent visitor in the last 60 days. However, tips may also reveal places where a user has been, since when posting tips users are often sharing experiences. [8] Finally, dones may also provide some evidence about a user's home location, although perhaps not as strong as tips and mayorships. Our conjecture is that users often mark as done tips about physical places where they have been to or intend to go soon. We note however that, despite intuitive, the aforementioned assumption is not guaranteed to hold for all users. As discussed in Temporal and Spatial Analyses section, 10% of the users in our dataset have a maximum displacement of at least 6,000 kilometers between consecutive tips and dones.

As a first step to address this question, we consider a simple approach that takes the most popular location among the attributes (mayorships, tips and/or dones) of a user as her home location, using a majority voting scheme. We note that more sophisticated methods could be applied such as classification algorithms (e.g., k-nearest neighbor) and other machine learning techniques [8, 11, 1]. Instead, we chose a simple majority voting approach as it allows us to assess the potential for effective inferences of this type in Foursquare.

We consider seven inference models which differ in terms of the attributes used for inference. The *Mayorship* model uses only the locations of the mayorships to infer the user's home location. Similarly, the *Tip* and *Done* models use only locations of tips and of dones, respectively. The *Mayorship+Tip*, *Mayorship+Done*, *Tip+Done* models use information from only two attributes, whereas the *All* model takes all three attributes jointly. By comparing alternative models, we are able to assess the potential of each attribute as source of inference. Moreover, recall that, as discussed in Mayorships, Tips and Dones section, there are non-negligible numbers of users that only have one or two of the attributes. Thus, the combination of multiple attributes may enable the inference for a larger user population. The models are here used mainly to infer the user's home city, although we also consider inferences about the user's home state and country.

To evaluate the effectiveness of each model, we take the information provided in the user's home city attribute as ground truth. Although users are free to enter whatever they want in this attribute, we found that the majority of Foursquare users do enter valid locations (see Table 1). To evaluate our in-

[8]Although users may post tips at unknown venues to, for instance, inquire about driving directions, operation time, or parking conditions, we believe that this does not occur very often.

Table 3. Home Location Inference.

Classes Distribution									
	Home City			**Home State**			**Home Country**		
Features	Class 0	Class 1	Class 2	Class 0	Class 1	Class 2	Class 0	Class 1	Class 2
Mayorship	727,179	847,876	239,129	707,953	913,166	110,110	727,179	1,053,703	33,302
Tip	725,073	671,576	192,781	702,583	727,219	99,672	725,073	835,532	28,825
Done	546,815	541,795	106,297	524,137	561,165	55,115	546,815	630,937	17,155
Mayorship+Tip	898,293	1,322,214	300,831	878,578	1,398,351	146,526	898,293	1,581,654	41,391
Mayorship+Done	825,009	1,213,917	270,974	805,029	1,278,784	130,439	825,009	1,447,581	37,310
Tip+Done	831,759	1,038,268	223,093	807,091	1,089,638	116,549	831,759	1,228,043	33,318
All	939,888	1,573,471	310,045	919,938	1,643,825	153,955	939,888	1,840,850	42,666
Accuracy									
	Home City			**Home State**			**Home Country**		
Features	Class 0	Class 1	Total	Class 0	Class 1	Total	Class 0	Class 1	Total
Mayorship	**51.61%**	**67.41%**	60.12%	**71.27%**	**80.92%**	**76.70%**	89.79%	92.92%	91.64%
Tip	51.52%	67.29%	59.11%	70.29%	80.59%	75.53%	**90.12%**	**93.67%**	**92.02%**
Done	50.09%	61.74%	55.89%	70.16%	78.38%	74.41%	89.12%	92.38%	90.87%
Mayorship+Tip	51.57%	66.24%	**60.31%**	70.21%	80.27%	76.39%	89.71%	93.13%	91.89%
Mayorship+Done	51.05%	65.27%	59.51%	70.01%	79.89%	76.07%	89.18%	92.78%	91.47%
Tip+Done	51.18%	64.16%	58.38%	69.76%	79.28%	75.23%	89.52%	93.04%	91.62%
All	51.46%	64.86%	59.85%	69.74%	79.53%	76.02%	89.29%	92.89%	91.67%

ferences, we consider only users whose home city attributes contain valid locations at the city level or at a finer granularity, as validated by *Yahoo! PlaceFinder*.

In our evaluation, we group users into three classes. *Class 0* consists of users who have a single activity, either a mayorship, a tip or a done. In this case, the unique choice is to set the user's home location equal to that of her activity. *Class 1* consists of users who have multiple activities with a predominant location across them. For these users, the inferred location matches the most often location of their activities. *Class 2*, in turn, consists of users with multiple activities in which there is no single location that stands out (i.e., there are ties). Our current inference approach cannot be applied to *Class 2* users.

Thus, we evaluate the proposed models by assessing their accuracy on users of both *Class 0* and *Class 1*. The accuracy corresponds to the percentage of correctly inferred locations out of all users of each class. Moreover, we also report the overall accuracy of each model, considering all users that are eligible for inference by the given model (i.e., users who have the required attributes).

Experimental Results

In this section, we present the experimental evaluation of our inference models. We start by discussing the results for inferring a user's home city, our main focus, discussing the inference of home state and country later in this section.

Table 3 shows, for each inference model, the number of users eligible for inference (i.e., users that have the required attribute) in each class (top of the table). It also shows the accuracy of the model for users in classes 0 and 1 as well as the overall accuracy considering all eligible users (bottom). We start by noting that the vast majority of the eligible users (87%- 91%) are in classes 0 and 1. Thus, for most users, either they have a single activity (33-45%) or they have multiple activities with a predominant location, and thus their home city can be inferred by our approach.

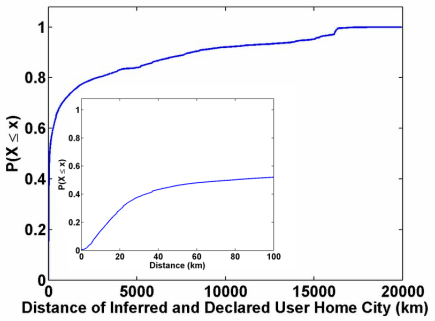


Figure 7. Cumulative Distribution of Distances Between Inferred and Declared User Home City.

We find that the models produce only marginally different results, in terms of accuracy, both per class and overall. As expected, mayorships are the best single attribute to infer home location, although, perhaps surprisingly, tips are only marginally worse. Dones, in turn, produce the worst results among the three attributes, when used in isolation. The combination of attributes does hurt the accuracy, in comparison with the *Mayorship* model, in most cases (*Mayorship+Tip* being the exception), possibly because tips and dones add some noise. However, note that, despite a somewhat lower accuracy, these combined models actually cover a much larger user population. For instance, the *Mayorship* model can only be applied to 1,814,184 users, whereas the *All* model is applicable to 2,823,404. Thus, considering the actual number of users for which each model was able to correctly predict the home city, we found that the best model was *All* (1,504,262 correct inferences) followed by *Mayorship+Tip* (1,339,152 correct inferences).

To better understand the models' errors, we computed for each incorrect inference the distance between the inferred city given by the *All* model and the declared user home city. Figure 7 shows the distribution of these distances. We found that around 46% of the distances are under 50 kilometers, which is a reasonable distance between neighboring (twin)

cities. Thus, combining these results with the correct inferences produced by our model, we find that we can correctly infer the city of around 78% of the users within 50 kilometers of distance.

We now turn our attention to the inference of a user's home state, whose results are also shown in Table 3. We note that, in comparison with the home city inference, all models improved for home state inference, reaching an overall accuracy around 75%. Once again, mayorships arise as the single attribute that produces the highest accuracy, for home state inference, followed by tips and dones. Nevertheless all models lead to very similar accuracies, both per class and overall. Thus, once again, due to the larger user coverage, the *All* model is able to correctly infer the home state of the largest number of users (1,948,851).

Finally, we also evaluate the models to infer a user's home country as a complementary analysis to validate our key assumption that users tend to have mayorships, tips and dones close to where he lives. As expected, Table 3 shows that all models achieve accuracies above 90% for home country inference. Unlike in the previous two cases, despite the great similarities in the results, the *Tip* model is the single attribute model that produces the best accuracy, followed by *Mayorship* and *Done*. The combined models produce very similar results, with *All* producing the largest number of correct inferences (2,549,177).

Our study presented satisfactory results in predicting user home location. Thus, an interesting implication of our work is that even the mispredictions may highlight some implicit user behavior in terms of mobility. At the city level, for instance, we observed some users that live nearby the inferred cities, which may indicate that they probably live in one place and move frequently to another. At the state level, the lower but non-negligible fraction of errors indicates that there are some users that have interstate mobility. Moreover, the inference of the home state may help disambiguate home cities, such as the case of Springfield. Finally, at the country level, we observed that there is a high concentration of the activities considered (mayorships, tips and dones) in the declared user home location. This can be verified by the higher accuracy that we obtained in our models. However, inference errors are still possible since some users may have his current home location outdated (e.g., a user who has just moved to another country) or may travel a lot around the world, or may even have a significant place-based identity with some city of another country (as discussed in [8]).

CONCLUSIONS AND FUTURE WORK

In this paper, we address the problem of privacy inference using publicly available features in Foursquare. Using a model that takes into account the majority of places where the user have interacted through mayorships, tips or dones, we are able to infer with high accuracy where the user current lives or his home location (city, state or country).

As future work, we plan to analyze the impact of differentiating features, e.g. giving weights, in the accuracy of our model. Also, we can explore more sophisticated machine learning approaches in attempt to increase our inference accuracy. Moreover, we plan to investigate other types of information that can be inferred using the same attributes.

Acknowledgements

This research is partially funded by the Brazilian National Institute of Science and Technology for Web Research (MCT/CNPq/INCT Web Grant Number 573871/2008-6), and the authors' individual grants from CNPq, CAPES and Fapemig.

REFERENCES

1. Z. Cheng, J. Caverlee, and K. Lee. You are Where You Tweet: a Content-based Approach to Geo-locating Twitter Users. In *Proc CIKM'10*.
2. Z. Cheng, J. Caverlee, K. Lee, and D. Sui. Exploring Millions of Footprints in Location Sharing Services. In *Proc AAAI ICWSM'11*.
3. E. Cho, S. Myers, and J. Leskovec. Friendship and Mobility: User Movement in Location-based Social Networks. In *Proc ACM SIGKDD'11*.
4. M. Choudhury, H. Sundaram, A. John, D. Seligmann, and A. Kelliher. "Birds of a Feather": Does User Homophily Impact Information Diffusion in Social Media? *CoRR'10*.
5. J. Cranshaw, R. Schwartz, J. Hong, and N. Sadeh. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. *ICWSM'12*.
6. C. Davis Jr., G. Pappa, D. Oliveira, and F. Arcanjo. Inferring the Location of Twitter Messages Based on User Relationships. *Transactions in GIS*, 15(6):735–751, 2011.
7. R. Gross and A. Acquisti. Information Revelation and Privacy in Online Social Networks. In *Proc WPES'05*.
8. B. Hecht, L. Hong, B. Suh, and E. Chi. Tweets from Justin Bieber's Heart: the Dynamics of the Location Field in User Profiles. In *Proc CHI'11*.
9. I.-F. Lam, K.-T. Chen, and L.-J. Chen. Involuntary Information Leakage in Social Network Services. In *Proc of IWSEC'08*.
10. N. Li and G. Chen. Sharing Location in Online Social Networks. *IEEE Network*, 2010.
11. J. Mahmud, J. Nichols, and C. Drews. Where Is This Tweet From? Inferring Home Locations of Twitter Users . In *Proc AAAI ICWSM'12*.
12. A. Mislove, B. Viswanath, K. Gummadi, and P. Druschel. You are Who You Know: Inferring User Profiles in Online Social Networks. In *Proc WSDM'10*.
13. S. M. News. http://www.socialmedianews.com.au/foursquare-reaches-15-million-users/.
14. A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An Empirical Study of Geographic User Activity Patterns in Foursquare. In *Proc ICWSM'11*.
15. T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors. In *Proc WWW'10*.
16. K. Tang, J. Lin, J. Hong, D. Siewiorek, and N. Sadeh. Rethinking Location Sharing: Exploring the Implications of Social-Driven vs. Purpose-Driven Location Sharing. In *Proc UBICOMP '10*.
17. M. Vasconcelos, S. Ricci, J. Almeida, F. Benevenuto, and V. Almeida. Tips, Dones and ToDos: Uncovering User Profiles in Foursquare. WSDM '12.
18. C. Vicente, D. Freni, C. Bettini, and C. Jensen. Location-Related Privacy in Geo-Social Networks. *IEEE Internet Computing*, 2011.
19. T. Vgele, C. Schlieder, and C. Schlieder. Spatially- Aware Information Retrieval with Graph-Based Qualitative Reference Models. In *Proc FLAIRS'03*.
20. Y. Zheng. Location-based social networks: Users. In Y. Zheng and X. Zhou, editors, *Computing with Spatial Trajectories*, pages 243–276. Springer, 2011.
21. D. Zwillinger and S. Kokoska. *CRC Standard Probability and Statistics Tables and Formulae*. Chapman & Hall, 2000.