Beware of What You Share: Inferring Home Location in Social Networks

Tatiana Pontes*, Gabriel Magno*, Marisa Vasconcelos*, Aditi Gupta†,
Jussara Almeida*, Ponnurangam Kumaraguru†, Virgilio Almeida*

*Universidade Federal de Minas Gerais, Brazil
{tpontes,magno,marisav,jussara,virgilio}@dcc.ufmg.br

†Indraprastha Institute of Information Technology, India
{aditig,pk}@iiitd.ac.in

***Abstract*—In recent years, social media users are voluntarily making large volume of personal data available on the social networks. Such data (e.g., professional associations) can create opportunities for users to strengthen their social and professional ties. However, the same data can also be used against the user for viral marketing and other unsolicited purposes. The invasion of privacy occurs due to privacy unawareness and carelessness of making information publicly available. In this paper, we perform a large-scale inference study in three of the currently most popular social networks: Foursquare, Google+ and Twitter. Our work focuses on inferring a user's home location, which may be a private attribute, for many users. We analyze whether a simple method can be used to infer the user home location using publicly available attributes and also the geographic information associated with locatable friends. We find that it is possible to infer the user home city with a high accuracy, around 67%, 72% and 82% of the cases in Foursquare, Google+ and Twitter, respectively. We also apply a finer-grained inference that reveals the geographic coordinates of the residence of a selected group of users in our datasets, achieving approximately up to 60% of accuracy within a radius of six kilometers.**

***Keywords*-Location; Privacy; Social Networks; Location Inference; Foursquare; Google+; Twitter**

I. Introduction

Online Social Networks (OSN) are one of the most popular web applications amongst Internet users. Initially, they were designed to connect close friends, but gradually new social networks were created with diverse purposes attracting users with different needs and reasons to sign up to this kind of system. Thereby, users are voluntarily making more personal information available such as their favorite places to visit, professional interests, personal views and reviews of company or service experiences. The availability of such data has several benefits like the development of personalization mechanisms and more effective recommendation strategies. Meanwhile, strengthening ties with the surrounding community maximizes users exposure for a varied audience spread in many systems. This potentially touches privacy concerns creating opportunities for unauthorized usage of user data.

Currently, Location-Based Social Networks (LBSN) have created new means for online interaction based entirely on the geographic location of their registered users allowing them to associate this kind of data with the shared data, facility which is being embedded also in OSNs. Disclosing individual data associated with location information could be even more invasive [20]. The collation of public location based attributes of a user aggregated over time may reveal her behavioral patterns and habits, emphasizing her preferences. Despite the privacy threats of sharing location, this is arising as a common behavior among users in Foursquare, which is currently the most popular LBSN, and even on the traditional OSNs, such as Google+ and Twitter.

Motivated by the possible privacy breaches due to the increased sharing of location information in social networks, here we perform a large-scale study on inferring the user home location in three of the currently most famous systems, namely Foursquare, Google+ and Twitter. Foursquare is a LBSN geared towards sharing of the instant location of users through check ins, which are converted in mayorships – title given to the most frequently visitor of a place (venue). Users may also leave notes (tips) about their experiences or impressions at specific venues, and also mark some previously posted tip with a sign of approval (like). These three types of information (mayorships, tips and likes) are public and are associated with the location (geographic coordinates) of a place. Google+ and Twitter present other variations in the way of sharing geographic data. A Google+ user can make public her home address, and also the institutions and companies where she has studied or worked so far, while on Twitter the tweets can be tagged with geographic coordinates, revealing where the user was when they were posted. Apart from these location shared data, the user profiles in all three systems also contain a home location attribute which is supposed to present the city where the user currently lives – in Google+, this attribute may contain a list of more than one place.

Our study consists of three main steps. First, we collected the public geographic information provided by users in the considered systems, building datasets containing millions of users. Second, we perform a characterization of these pieces of information that are potentially relevant to infer the user home location. Finally, we propose and evaluate models that use those attributes to infer the city where the user lives and his exact residence location. We correctly infer the user home city with a high accuracy, in around 67%, 72% and 82% in Foursquare, Google+ and Twitter, respectively. We also achieve successful inferences for the exact home residence of a selected group of Foursquare and Twitter users within a radius of six kilometers with about 60% of accuracy. For Google+, the equivalent accuracy is comparatively lower due to the nature of the attributes of the system. From our results, we conclude that sharing location information on Foursquare and Twitter may lead to critical privacy leak by revealing the home residence location of users.

Rest of this paper is organized as follows: Section II discusses related work while Section III describes our datasets. Section IV presents a characterization of the pieces of information in each social network. Section V discusses the inference strategies and main experimental results. Section VI concludes our findings, pointing out possible directions for future work.

II. Related Work

The increasing share of personal information through a diverse range of social networks with different purposes is raising concerns about privacy related issues. Some studies have shown that private data can be easily disclosed by the collation of the set of user attributes in a system [8]. Thus, these possible inferences, basically, allow that explicit information reveals implicit data of the user making him more vulnerable and exposed [10]. Choudhury *et al.* argued that user homophily does influence the information diffusion in social networks suggesting that users with similar preferences tend to be friends, which opens a privacy breach to explore users through their social network [5]. Similar concepts are addressed in other studies which show that personal interests of users can be inferred from friends [14], and also the profile attributes of a user may be revealed through friends by analyzing their tagged-photos [16].

Recently, several studies have focused on investigating the user geographic information to understand aspects related to human mobility patterns [3], [4], [15], city urban development [6], nature event detection [18] and also the impact on privacy from users' location sharing [11]. A few studies have tried to estimate the location of a user using his attributes associated with any geographical information. In [7], authors proposed a model for inferring the home location of Twitter users through their friends location assuming that the user social network usually consists of people who are likely to live nearby. As in this work, we are also able to obtain satisfactory results which reinforce this premise. Also, the lack of geographic-based features used by the Twitter users motivated the design of inference models based on tweet textual content [2], [9], [13]. Cheng *et al.* created a model based on the common vocabulary of users from the same geographical region. Authors in [9] used machine learning strategies to infer the user home state and country exploiting the textual content of her tweets. Finally, Mahmud *et al.* also used machine learning techniques considering the content of tweets to infer the user home city, state and time zone [13].

In this paper, we extend the work done by [17], which addresses concerns about privacy violation associated with the inference of the users' home location in Foursquare. Unlike the previous related studies, we here propose and evaluate different inference models for three social networks, namely Foursquare, Google+ and Twitter. We exploit large datasets with millions of users aiming to reveal the potential of public geographic-related attributes in disclosing the user home location. Moreover, we experiment with the home location of the user's friends as inference attribute. The inference applied is performed in the spatial granularity of cities and also in the level of geographic coordinate of the user residence. To the best of our knowledge, this is the first privacy work conducted in three different systems for location inference based on the discovery of the exact user residence location.

III. Location-Aware Online Social Networks

In this section, we briefly review the social networks and introduce the datasets used in our inference study. That is, we present the main system components and summarize the datasets collected from Foursquare (Section III-A), Google+ (Section III-B) and Twitter (Section III-C).

A. Foursquare

Foursquare is currently the most popular LBSN providing support to location sharing with friends through *check ins*. Check ins are performed only via devices with GPS and are associated with places (*venues*), which represent real locations previously registered in the system – such as restaurants, monuments or residences.[1] The larger the number of check ins a user does, the more incentives she may earn to continue sharing. By incentives we mean, for instance, *mayorships*, which is a title given for the most frequent visitor of a venue in the last 60 days.

Although Foursquare was initially created with the primary intention of promoting a game between users competing for check ins as well as mayorships, it also includes attributes (*tips* and *likes*) that favor the recommendation of places among users. Tips are comments left by users on specific venues which reflect the users' experiences and opinions about some aspect of visited places (e.g., the quality of service or availability of parking space in a restaurant or even instructions about how to find the place). A like, on the other hand, is a sign of agreement with the content of a previously posted tip.

In Foursquare, the location information available in public users' attributes are the home city as well as the history of mayorships, tips and likes which are associated to venues which also have a public location data. Our dataset, crawled from August to October 2011 through the system API,[2] comprises of attributes aggregating location information associated with 13,570,060 users and 15,898,484 different venues collected. The user home city is an optional open text field limited in 100 characters where the user is supposed to write the city where he lives. For venues, the location must be defined filling the open text fields, namely city and address (limited in 30 and 127 characters, respectively), and also

[1] Residence is a venue category related to real homes. Their coordinates are ommited in the venue's page, but are accessible via Foursquare API.

[2] The Foursquare dataset was used to characterize the use of tips and likes [19] and also to analyze the privacy of users [17].

setting a pin in a map. Note that the system does not provide any automatic tool to enforce users and venues owners to write valid information in their location fields. In summary, the set of attributes which we explore in this paper are the user home city, the friends home city, and the city of the venues related to the user history of mayorships achieved, tips posted and likes given.[3] Our entire dataset consists of 15,149,981 mayorships, 10,618,411 tips and 9,989,325 likes.

B. Google+

Google+ is an online social network owned by Google which focuses on information sharing through labeled groups of users. These groups, called *circles*, represent a specified subset of the followers of a user with whom she may share or receive information. A user may, for instance, manage "family", "colleagues", and "alumni" circles, filtering the visibility of his own shared content and also avoid receiving unwanted messages as feed. A user may set the level of visibility of each piece of content she shares by choosing which of her circles are allowed to see it. The relationship between users are of follow and be followed, thus friendships are derived from these relations considering the reciprocal links between users in the same circles.

Our Google+ dataset consists of publicly available user profile information collected from November to December 2011.[4] In total, we crawled 27,556,390 profile pages through HTTP requests to the system. For each user profile, we collected the circles to get the complete list of friends of the user as well as the public user data. We focus on the attributes associated with location such as the user *address*, his list of *places lived* including all cities he has lived in, over time, the set of *education* institutions where she has studied and also the *employment* attribute which lists the companies where he has worked. Out of all collected users, 7,371,461 defined at least one place where they lived, 5,162 provided an address information and 7,471,191 and 5,917,609 filled the education and employment with at least one institution or company, respectively. All these attributes are optional open text fields which can be made public or private. Only the education and employment attributes are supported by the system automatic filling tool which helps users to complete the field, but also allows them to include whatever they want in this attribute.

C. Twitter

Twitter is an online social network as well as a microblogging which has gained popularity over the last few years as a major information and opinion sharing medium enabling users to participate in the system with text messages of up to 140 characters, known as *tweets*. These tweets may also be geographically tagged being associated with the instant position of GPS enabled mobile devices in the moment of the tweeting activity. As our focus on this paper is on exploring location information, we can say that on Twitter there are two primary ways in which it can be shared: firstly, as a user profile attribute called *location* which is an optional and public open text field limited in 30 characters where the user is supposed to write his home city name; and secondly, as a geographic coordinate associated with the *geographically tagged tweets* of a user which can be made as private by her.

The data was collected using the Twitter streaming API,[5] using the methodology discussed in [1]. We crawled tweets obtained through a filter that considers only the ones related to the most popular topics, thus presenting some word query term. We crawled around 120,331,140 tweets posted by 19,684,469 unique users from April to June 2012. From these, we observed that only about 716,681 tweets (0.5%) posted by 295,307 unique users were geographically tagged.

IV. Dataset Properties

In this section, we first standardized the location information associated with the geographically referenced attributes in our datasets and analyzed the "quality" of these data in terms of the level of spatial granularity (Section IV-A). Then, in Section IV-B, we characterized those attributes, which are used in our proposed location inference models, assessing their usage in each analyzed system.

A. Geographically Referenced Information

In all three datasets, the considered attributes are supposed to be associated with valid geographic information. However, in most analyzed attributes (except the coordinates associated with the geographically tagged tweets in Twitter), this information is supplied in an open text field, which means the users can write whatever they want without any automatic verification. Thus, a lot of noisy due to invalid locations, misspelling or even non sense words may appear.

To filter text that does not correspond to a valid location information, we used the *Yahoo! Places Finder* geo-coding API.[6] We have also used the *Yahoo!* results to standardize some location names, disambiguating all the possible variations found for a place (e.g., NY, New York City, etc). Valid location responses in *Yahoo!* include among the geographic coordinates, city, state and country names and a "quality" indicator which is an integer value between 0 – 99 that represents the best spatial granularity matched for the correspondent query. For instance, for a query like "New York", the *Yahoo!* response would present a "quality" of 40 indicating that it is in the city level, the best possible matching for this query.

Table I provides the distribution of the geographic information (GI) of all considered attributes in each dataset. We present, for each attribute, the percentage of it that correspond to valid geographic information (real location), non-geographic information (e.g., emails, phrases) or no information declared (empty). The valid geographic information can be unambiguous (UGI) or ambiguous (AGI), once there are some city names which may refer to multiple cities in the Earth

[3] Check ins are not considered here as they are a private attribute, and thus their geographic location are not accessible via Foursquare API.

[4] The dataset was used to analyze the social graph of Google+ and also privacy aspects related to the registered users in the system [12].

[5] https://dev.twitter.com/docs/streaming-api

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

TABLE I

Availability of Geographic Information (GI) in Various Attributes in our Datasets. All values presented are in percentage (%).

	Foursquare		Google+				Twitter	
Statistics	**User Home City**	**Venue City**	**Places Lived**	**Address**	**Education**	**Employment**	**User Location**	**Geo-tagged Tweet**
% valid UGI	95.35	55.45	61.85	0.01	52.95	34.52	73.28	100.00
% valid AGI	2.65	18.04	6.66	0.002	11.01	14.67	9.70	0.00
% non-GI	1.80	26.51	31.48	0.01	36.04	50.81	11.90	0.00
% empty	0.20	0.00	0.00	99.98	0.00	0.00	5.12	0.00

indistinctly, being *Yahoo!* unable to decide which is correct – e.g., “Springfield” is the name of ten different cities, only in The United States. For Foursquare, due to space constraints in the paper, we group tips, likes and mayorships as venue attributes, while users attributes correspond only to the home city field. Note that the vast majority of Foursquare users (98% of 13,570,060) provided valid home city locations, with only a tiny fraction leaving it blank (0.2%) or filling it with non-geographic information (1.8%). Moreover, 11.6 million venues have valid locations associated, although a substantial fraction of all venues have non-valid locations (around 26%) or valid but ambiguous location (18%). This large fraction of non-valid or ambiguous venue locations comes as a surprise, particularly considering that, unlike the user home city field, the venue location information is a mandatory attribute.

In comparison with Foursquare, the fraction of valid locations in our Google+ dataset is much lower for all considered attributes. Note that the fraction of users with valid locations in their places lived field is higher (61.85% of 27,556,390 users) than those with valid locations in the education and employment fields, possibly because many users fill those fields with institution or company names, which cannot be recognized by *Yahoo!*. Note also that only a tiny fraction of the users share their addresses (5,162 out of over 27 million users analyzed). Moreover, as Google+ users may opt not to publicly display their attributes, we cannot distinguish between private and empty fields. Thus, we here treat both as empty.

In Twitter, we see that all geographically tagged tweets in our dataset (716,681 in total) contain valid location information associated, which is represented by a valid latitude and longitude pair. Considering the set of users who tag their tweets, 94% provide location information in their profiles, being 82.9% valid and 11.9% not valid geographic information.

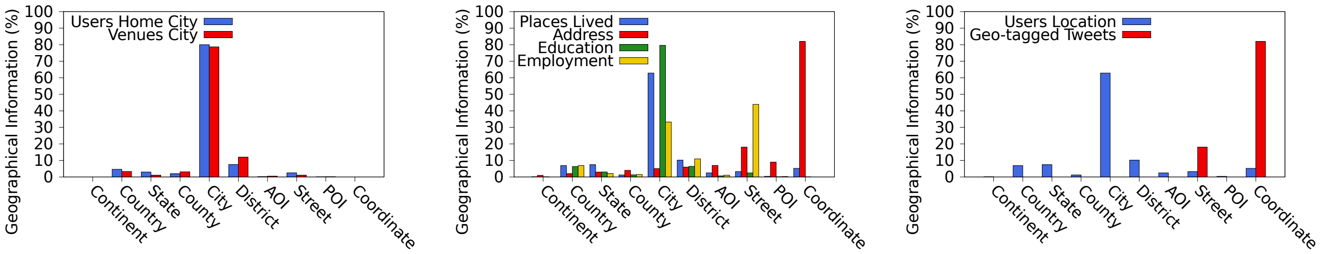
Next, we analyze the “quality” of the valid (and unambiguous) geographic information available in the datasets of the social networks addressed. The distributions of the “quality” of the information provided in the analyzed attributes are shown in Figure 1. According to Figure 1(a), the vast majority (80%) of Foursquare users and venues have location information at the city level. Only 9.62% and 7.54% of users and venues present coarser location granularities (e.g., at state or country levels), and the fractions with finer-grained positioning (from district to coordinate level) reach 10.36% and 13.76%, respectively. The same user behavior is observed in Google+ (Figure 1(b)) and Twitter (Figure 1(c)), where the majority of the users on these systems (79.63% and 62.54%, respectively) provide the home location information at the city level. However, for Google+ users, the location information associated with the education, employment and address attributes are more often provided at finer granularities, i.e., street level for employment and address, and Point Of Interest (POI) for education. Finally, the “quality” of the location provided in users’ tweets is either at the street (18.05%) or at the geographic coordinate (81.95%) levels. The availability of public finer-grained location information opens an opportunity for more specific inferences regarding user home location, such as user residence, as discussed in Section V-C.

B. *Attribute Characterization*

In the previous section, we analyzed the availability of valid and unambiguous geographic information as well as the “quality” of this information across all analyzed attributes. Now, we focus on the usage of these attributes and analyze their distributions across users in each dataset. We aim at assessing the potential of exploiting these attributes for inference purposes in terms of the fraction of users we would cover.

Starting with Foursquare, Figure 2(a) shows the cumulative distributions of the numbers of mayorships owned, tips posted and likes given per user. Clearly, all three distributions are heavy tailed, since most users tend to have few mayorships (tips or likes), whereas a few users are very active considering these attributes. The curves are very similar and show that, for each attribute, 90% of the users considered have up to 10 mayorships (tips or likes). Nevertheless, we find that, out of all users in our dataset, almost 4.2 million (i.e. 30%) have at least one mayorship, tip or like, whereas about 890 thousands users have all these attributes. Moreover, 1 million users have only mayorships and about 670 and 367 thousands have only tips or likes, respectively. Thus, exploiting these attributes to infer home location seems to be promising as the required information is available for a large number of users.

In Google+ dataset, around 10.7 million of the users (39%) have defined at least one location in the places lived field, at least one education institution, one employment location, or provided any address information. In total, 1,878 users have all the above fields filled. However, excluding the address information, 2.9 million users (11% of our entire dataset) have at least one location in each of the other three fields. Also, about 1.6 million users have only filled the places lived attribute, whereas 1.4 million and 745 thousand have only education and employment information, respectively. Once again, we find that the cumulative distributions of these attributes across Google+ users, shown in Figure 2(b), are heavy tailed. The graphs show that only a small fraction of users has lists of attributes with sizes greater than 1, being around 6% for places lived, 2.5% for education and 1% for employment. Thus, as



(a) Foursquare

(b) Google+

(c) Twitter

Fig. 1. Quality of the Valid and Unambiguous Geographic Information.

can be seen, for all curves a tiny percentage of users filled the attributes with more than 10 places.

Now, looking into the Twitter dataset, we can observe in Figure 2(c), the cumulative distribution of the geographically tagged tweets posted by users in the system. We can see that less than 5% of users (out of 295,307 which have at least one geographically tagged tweet) have shared more than 10 tweets with this location information associated, thus emphasizing a common behavior in the system of not tagging various tweets. This can be clearly observed in the beginning of the curve which shows that most of the users (62%) have posted only one tweet with geographic tag.

Finally, in addition to the aforementioned attributes, we also consider exploiting the location information associated with the friends of a user for the purpose of inferring her home location. Thus, we now analyze the distributions of the number of friends of users in our Foursquare and Google+ datasets. As our Twitter dataset does not contain information of who follows (or is followed by) each user, we did not consider this analysis for this particular dataset. The cumulative distributions of friends per user in Foursquare and Google+ datasets are presented in Figures 2(a) and 2(b), respectively. We note that the distributions are very skewed for each system, with about 98% of Foursquare users having up to 100 friends while only 2% reach almost 7,000 friends. Differently, for Google+, 18% of the users have at least two friends, whereas only 1% have 10 or more.

V. Inferring Location

In this section, we present our proposed models for inferring location for the three considered systems. We start by discussing our methodology for the experimental evaluation in Section V-A. Next, we discuss the results of applying our location inference models at both the city (Section V-B) and geographic coordinate levels (Section V-C).

A. Methodology

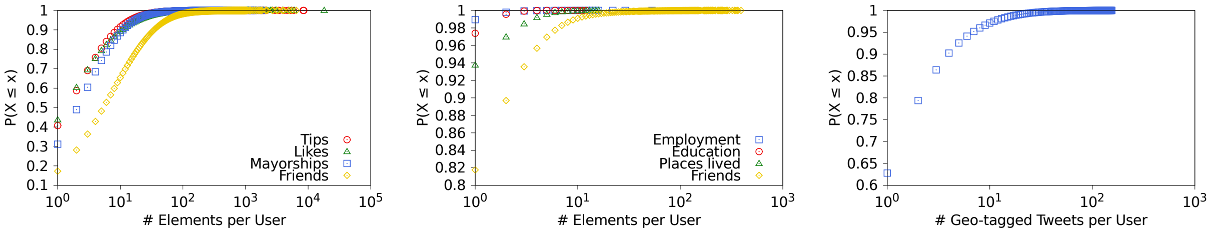
The location inference models proposed in this work consider only publicly available users' attributes for all of our three datasets. The key assumption here is that the nature of these attributes suggests that they are in some way associated to nearby locations from the user's home. In Foursquare, for instance, users can only achieve a title of mayor after frequently visiting a place, and we believe that also tips and likes reflect any kind of user experience while visiting the place. The institutions or companies where one has studied or worked, as well as the address declared in Google+ user profile indicate the places where the user has lived or currently lives. Associating a geographic information in Twitter posts is the same as gluing a sticker at each place one tweets, i.e. leaving footprints with it. Thus, the combination of all tweets' location may reveal the user's routine. Finally, previous work [7] uses the locations of friends to predict a user home location, based on the assumption that the most users friends tend to live in the same city. Building on these previous efforts, we here also explore the friends' locations in our inference models.

Our primary goal in this work is to apply inference models considering all those location attributes to verify if they are indeed associated with the same city which the user declare living. As a first step to address this problem, we here consider a simple majority voting scheme which assigns the most popular location among the user's attributes as her home location. In other words, the location associated with each available attribute is taken as a vote to a specific city. The city that gets the largest number of votes is inferred as the user home city. More intelligent methods such as machine learning techniques as well as classification algorithms (e.g. k-nearest neighbor) could also be applied.[7] Instead, we chose a simple majority voting approach as it allows us to assess the potential for effective inferences of this type in the analyzed systems. Indeed, our preliminary efforts show that this simple approach can be reasonably effective.

In our experiments, we group users into three classes that differ by the number of votes aggregated per city, *Class 0* consists of users who have only one vote (i.e., a single attribute with location information), thus allowing only a unique option to be assigned for the user home city. *Class 1* contains users who have multiple votes with a predominant location across them. The inferred location for these users matches the most often location in their votes. *Class 2*, in turn, consists of users with multiple votes 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.

The results of our experimental evaluation are assessed using two metrics which measure the effectiveness of the proposed models. *Accuracy* is the fraction of correct inferences applied on users of classes 0 and/or 1. Moreover, we report the *amount of users covered* which corresponds to the number

[7] We plan to look at these advanced techniques as part of our future work.



(a) Foursquare

(b) Google+

(c) Twitter

Fig. 2. Cumulative Distribution of the Location-based Attributes in each of the Three Media (log scale in the x-axis).

of users for which we could infer a home location (users in classes 0 and 1), as our intention is to apply inferences for a great percentage of users.

B. City-Level Inference

In this section, we present our experimental evaluation for home location inference at city level. In Section V-B1, we present the proposed models for each dataset, whereas, in Section V-B2, we evaluate them according to both inference accuracy and user coverage.

1) Inference Models: We propose different inference models for each dataset, as they exploit attributes that are specific to each analyzed system. We consider models that exploit each attribute in isolation as well as combinations of attributes. The ground-truth of the inference models in the city level were set as the declared user home city in Foursquare, the city presented in the list of places lived of the Google+ users[8] and the location attribute in the Twitter users profile.

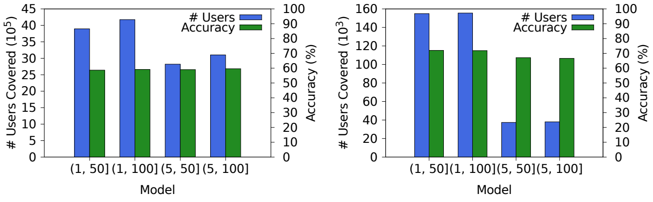
We build four single-attribute models for Foursquare, referred to as *Mayorship*, *Tip*, *Like* and *Friend* models. For Google+, we also build a *Friend* model as well as an *Education* and an *Employment* model, all of them based on a single attribute. For Twitter, the only attribute used in the inference task is the set of geographically referenced tweets posted by the user, which we refer as *Geo-tagged Tweet* model. For both Foursquare and Google+, we also combine different attributes to build alternative models, aiming at assessing the potential of these attributes to improve the accuracy of the isolated models and increase the number of users covered in the inference. We consider all possible combinations of groups of attributes, but due to space constraints, we present results only for the combination of all attributes, here referred as *All* model.

We also experiment with a refinement for the *Friend* models, which consists of filtering users with very few (i.e., less than k_{min}) or too many (i.e., more than k_{max}) friends out of the inference process. This refinement, originally proposed in [7], is motivated by the conjecture that these users may represent noise to the inference as users with very few friends lack enough evidence from which to build the inference, whereas users with too many friends probably do not have strong relationships with all their friends. We evaluate the benefits from this refinement for various values of k_{min} and k_{max}.

2) Results: The experimental results for each of our datasets regarding all analyzed models are presented in Table II. For each proposed model, the table shows the number of eligible users for the inference task, i.e., the number of users who have at least one of the attributes required by the specific model. It also shows the distribution of these users across the three previously defined classes, along with the model accuracy for users in *Class 0* and *Class 1*, as well as an average accuracy. Table II shows results for the *Friend* method *without* any refinement.

We start by noting that the number of users eligible for inference is much larger in Foursquare than in the other two systems, exceeding 7.1 million for the *All* model. Moreover, we find that the vast majority of these users (63%-100%) are in classes 0 and 1, and thus are covered by our proposed inference models.

Comparing to single-attribute models, we find that, mayorships are the best single attribute to infer home location in Foursquare, which is intuitive as they are derived from frequent check ins (Section III-A), and thus provide a strong piece of evidence regarding a user's home city. Surprisingly, tips are only marginally worse than mayorships, whereas likes and friends are clearly weaker sources for inference. In Google+, in contrast, the list of friends is the best single attribute, probably because people often live and study/work in different cities, thus making education and employment less reliable attributes.



(a) Foursquare

(b) Google+

Fig. 3. Home City Inference for the Refined *Friend* Model.

We also find that the *All* model accuracy is affected by the combination of multiple attributes, as less accurate attributes introduce noise to the inference.[9] The detrimental impact on accuracy is particularly strong on Google+, where there is great variability in the accuracy of single-attribute models. As consequence, average accuracy drops from around 51%

[8] The Google+ users with more than one location in the places lived attribute were disregarded as the model ground-truth can not be dubious.

[9] In Foursquare this is not observed as the model that takes mayorships and tips into account has an average accuracy of 60.31%, while covering at least 24% more users in comparison with the models with attributes in isolation.

TABLE II
SUMMARY OF THE RESULTS OBTAINED FOR THE INFERENCE MODELS FOR HOME CITY INFERENCE.

			Classes Distribution			**Accuracy**		
Dataset	**Inference Models**	**# Eligible Users**	Class 0	Class 1	Class 2	Class 0	Class 1	Total
Foursquare	*Mayorship*	1,814,184	40.08%	46.74%	13.18%	**51.61%**	**67.41%**	**60.12%**
	Tip	1,589,430	45.62%	42.25%	12.13%	51.52%	67.29%	59.11%
	Like	1,194,907	45.76%	45.34%	8.90%	50.09%	61.74%	55.89%
	Friend (no refinement)	6,973,727	17.27%	61.56%	21.17%	33.03%	59.26%	53.51%
	All	**7,153,078**	16.69%	64.15%	19.16%	35.28%	61.03%	55.72%
Google+	*Education*	1,171,456	88.27%	1.30%	10.43%	21.17%	48.80%	21.57%
	Employment	619,265	92.41%	0.46%	7.13%	7.56%	22.27%	7.64%
	Friend (no refinement)	599,649	52.00%	25.97%	22.03%	**40.43%**	**71.82%**	**50.89%**
	All	**1,538,227**	46.71%	16.02%	37.27%	17.37%	67.71%	30.22%
Twitter	*Geo-tagged Tweet*	**196,653**	89.66%	10.34%	0.00%	**82.50%**	**79.17%**	**82.16%**

to only 30%. Nevertheless, these models achieve the largest user coverage, with about 1.5 and 7.1 million eligible users in Google+ and Foursquare, respectively. Thus, there is a clear trade-off between both metrics. Indeed, note that, despite a somewhat lower accuracy, these combined models make correct inferences for a much larger user population: about 3.2 million users in Foursquare and 291 thousand in Google+.

Similarly, we find that, in terms of accuracy, only the results for Twitter are far better than the best results for Foursquare, which in turn exceed those for Google+. However, the fraction of all users collected from Twitter that are eligible for inference (1%) is much smaller than the fractions in Foursquare (52.7%) and Google+ (5.5%).

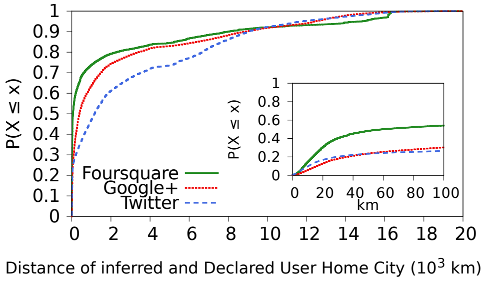


Fig. 4. Distance Between Real and Inferred User Home City.

We now discuss the impact of refining the *Friend* model by filtering users that have up to k_{min} or more than k_{max} friends out of the inference. Figure 3 shows the total accuracy which considers the inferences for users in classes 0 and 1 and the number of users covered by the refined *Friend* model for various values of k_{min} and k_{max}, specified in the x-axis of the graphs as pairs (k_{min}, k_{max}) of an interval. Comparing the results with those shown in Table II, we see that the refinement improves model accuracy, particularly for Google+, where the gains reach 21%. They come at the cost of a reduced user coverage. In Foursquare, the accuracy improvements are around the same for all values of k_{min} and k_{max} tested (around 6%). However, the number of users covered drops greatly if we increase k_{min} from 1 to 5 (a decrease of as much as 27.5%). The impact of increasing k_{max} from 50 to 100, on the other hand, is smaller, as the number of users with a list of friends between 50 and 100 is very small (representing only around 6% and 9% of the users covered in the configurations of (k_{min},k_{max}) of [1,100) and [5,100), respectively). The results for Google+ are quantitatively similar, although the reduction in user coverage is more significant (up to 88.2%) as we increase k_{min} to 5.

To better understand the errors in the models which led us to make erroneous inferences for users, we computed for each incorrect inference, the spatial distance between the inferred city and the one set in the ground-truth. The cumulative distribution of these distances for our most accurate models to each dataset (which their total accuracies are in bold in Table II) are presented in Figure 4. Observe that Figure 4 corresponds only to the incorrect inferences and the inner graph is, basically, a zoom in the outer graph. It shows that around 46% of the distances in Foursquare, and also 27% in Google+ and Twitter are under 50 kilometers which is a reasonable distance between neighboring cities. Thus, combining these results with the correct inferences produced by our models, we can make correct inferences in a radius of 50 kilometers with accuracies that achieve 78.5% in the Foursquare, 64.2% for Google+ whereas in Twitter we have 87%. As we can see, the Google+ results were not as good as the ones obtained from the other social networks. This is due to the nature of Google+ features which may be not associated with nearby places leading to less gains when increasing the threshold of the tolerance distance between inferred and real geographic coordinates of users home city.

As a final note, we point out that the fraction of users in *Class 2* is significant in both Foursquare and Google+. These users are not eligible for inference by our current models as they have no predominant location in the considered attributes. As a future work, we intend to investigate alternative approaches to address this kind of tied results.

C. Geographic Coordinate-Level Inference

In this Section, we present the experimental evaluation of the user residence inference in the geographic coordinate level. We describe the proposed inference models in Section V-C1 for each of our datasets while, in Section V-C2, we analyze and discuss the results achieved.

1) Inference Models: As in the city level inference, we propose different models for each specific social network addressed in this work. However, here, only one combined model was developed for each system considering all the listed attributes except the friends location.[10] As ground-truth, we set the user residence as the location associated with the venue of the Residence category where he was mayor in Foursquare. In Google+, we considered the address information presented in the user profile, and, at last, in Twitter, we considered the user location attribute with a "quality" value in the level of geographic coordinate.

We used two phases to infer the geographic coordinates of a user residence in each system. First, we apply the majority voting scheme to infer the user home city. After that, we compute the mean of the coordinates (pair with the mean latitude and longitude) of the attributes which are within the inferred city. Thus, to evaluate how good our inference is, we plot the cumulative distributions of the distances between the inferred coordinate and the one associated with the ground-truth - which represent the exact location of the user home.[11]

2) Results: In total, 1,272,919 users were eligible for the residence inference in Foursquare, 516 in Google+ and 10,140 in Twitter. By eligible, we mean users who have a ground-truth available and also some other attributes to be used by the inference model. In Figure 5, we show the cumulative distribution of the distances between real and inferred users residence location for all inferences made - the inner graph is a zoom in the outer one. For Twitter, we have 35.41% of the inferences with distances equal to zero (i.e., the proposed model inferred exactly the user residence location) and 73.67% are within a radius of 20 km, indicating that there are users tweeting close to their residences. By looking at Foursquare results, we had 52.73% of the inferences in a radius smaller than 5 km, and 77.27% less than a 20 km radius. Finally, for Google+, we are only able to infer the exact residence location of 5.23% of the users, which is expected, since we are using attributes of places where the user studied or worked. Thus, these higher distances errors for Google+ suggest that people may live and work in different cities.

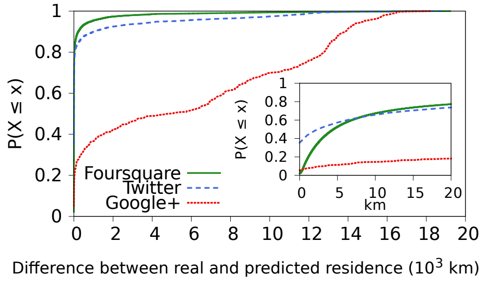


Fig. 5. Distance Between Real and Predicted User Residence.

VI. Conclusions

In this paper, we addressed the problem of privacy invasion using publicly shared attributes on three popular social networks: Foursquare, Google+ and Twitter. For each system, we considered models based on attributes of the medium to perform an inference about the users home city and also their residences location. Our results shown that it is possible to infer the user home city with a high accuracy, in around 67%, 72% and 82% of the cases in Foursquare, Google+ and Twitter, respectively. In the case of a finer-grained inference, our proposed models were able to compute the residence location of Foursquare and Twitter users within a radius of six kilometers with approximately 60% of accuracy. Though, the Google+ model presented a low accuracy of 10% for residence location inference. From our analysis, we conclude that location information sharing on Foursquare and Twitter can lead to critical privacy leak by revealing the residence location for many of its users.

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] A. Gupta, P. Kumaraguru. Credibility Ranking of Tweets During High Impact Events. In *Proc PSOSM'12*.

[2] 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*.

[3] Z. Cheng, J. Caverlee, K. Lee, and D. Sui. Exploring Millions of Footprints in Location Sharing Services. In *Proc AAAI ICWSM'11*.

[4] E. Cho, S. Myers, and J. Leskovec. Friendship and Mobility: User Movement in Location-based Social Networks. In *Proc ACM SIGKDD'11*.

[5] 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*.

[6] J. Cranshaw, R. Schwartz, J. Hong, and N. Sadeh. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. *ICWSM'12*.

[7] 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.

[8] R. Gross and A. Acquisti. Information Revelation and Privacy in Online Social Networks. In *Proc WPES'05*.

[9] 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*.

[10] I.-F. Lam, K.-T. Chen, and L.-J. Chen. Involuntary Information Leakage in Social Network Services. In *Proc IWSEC'08*.

[11] N. Li and G. Chen. Sharing Location in Online Social Networks. *IEEE Network*, 2010.

[12] G. Magno, G. Comarela, D. Trumper, M. Cha, and V. Almeida. New Kid on the Block: Exploring the Google+ Social Graph. In *Proc IMC'2012*.

[13] J. Mahmud, J. Nichols, and C. Drews. Where Is This Tweet From? Inferring Home Locations of Twitter Users . In *Proc AAAI ICWSM'12*.

[14] 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*.

[15] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An Empirical Study of Geographic User Activity Patterns in Foursquare. In *Proc ICWSM'11*.

[16] J. Pesce, D. Casas, G. Rauber, and V. Almeida. Privacy Attacks in Social Media Using Photo Tagging Networks: A Case Study with Facebook. In *Proc PSOSM'12*.

[17] T. Pontes, M. Vasconcelos, J. Almeida, P. Kumaraguru, and V. Almeida. We Know Where You Live: Privacy Characterization of Foursquare Behavior. In *Proc LBSN'12*.

[18] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors. In *Proc WWW'10*.

[19] M. Vasconcelos, S. Ricci, J. Almeida, F. Benevenuto, and V. Almeida. Tips, Dones and ToDos: Uncovering User Profiles in Foursquare. WSDM '12.

[20] C. Vicente, D. Freni, C. Bettini, and C. Jensen. Location-Related Privacy in Geo-Social Networks. *IEEE Internet Computing*, 2011.

[10] Since the friends location is defined as a default coordinate of their home city, we do not consider friends in our inference models for residence.

[11] In cases of Foursquare users with multiple mayorships in venues of the Residence category, we decided to report the lowest error.