It's the Way you Check-in: Identifying Users in Location-Based Social Networks

Luca Rossi
School of Computer Science
University of Birmingham, UK
l.rossi@cs.bham.ac.uk

Mirco Musolesi
School of Computer Science
University of Birmingham, UK
m.musolesi@cs.bham.ac.uk

ABSTRACT

In recent years, the rapid spread of smartphones has led to the increasing popularity of Location-Based Social Networks (LBSNs). Although a number of research studies and articles in the press have shown the dangers of exposing personal location data, the inherent nature of LBSNs encourages users to publish information about their current location (i.e., their *check-ins*). The same is true for the majority of the most popular social networking websites, which offer the possibility of associating the current location of users to their posts and photos. Moreover, some LBSNs, such as Foursquare, let users tag their friends in their check-ins, thus potentially releasing location information of individuals that have no control over the published data. This raises additional privacy concerns for the management of location information in LBSNs.

In this paper we propose and evaluate a series of techniques for the identification of users from their check-in data. More specifically, we first present two strategies according to which users are characterized by the spatio-temporal trajectory emerging from their check-ins over time and the frequency of visit to specific locations, respectively. In addition to these approaches, we also propose a hybrid strategy that is able to exploit both types of information. It is worth noting that these techniques can be applied to a more general class of problems where locations and social links of individuals are available in a given dataset. We evaluate our techniques by means of three real-world LBSNs datasets, demonstrating that a very limited amount of data points is sufficient to identify a user with a high degree of accuracy. For instance, we show that in some datasets we are able to classify more than 80% of the users correctly.

Categories and Subject Descriptors

K.4 [**Computers and Society**]: Public Policy Issues—*Privacy*; H.4 [**Information Storage and Retrieval**]: Online Information Services—*Data sharing, Web-based services*

Keywords

Location-based social networks; User identification; Privacy

COSN'14, October 1–2, 2014, Dublin, Ireland.

ACM 978-1-4503-3198-2/14/10 ...$15.00.
http://dx.doi.org/10.1145/2660460.2660485.

1. INTRODUCTION

With the proliferation of GPS and Internet enabled smartphones over the last years, Location-Based Social Networks (LBSNs) have been increasingly popular and have attracted millions of users. Examples of LBSNs include BrightKite[1], Gowalla[2], Facebook Places [3] and Foursquare[4]. While BrightKite and Gowalla have been discontinued, Foursquare is now one of the most popular and widely used LBSNs with nearly 30 million users and over 3 billion check-ins.

These systems are based on the concept of *check-in*: a user can register in a certain location and share this information with his/her friends with the possibility of leaving recommendations and comments about shops, restaurants and so on. However, a great deal of research has highlighted the dangers of exposing personal location information [2, 5, 20]. In particular, the problem of protecting privacy in LBSNs has also been the subject of several studies, such as [4, 17, 19, 33, 7, 22, 28, 29]. The social nature of LBSNs inevitably introduces new concerns, as users are encouraged to disseminate location information on the network [31]. Moreover, as noted by Ruiz et al., the practice of tagging users can lead to the release of location information about other individuals that have no control over the published data [31]. For instance, in August 2012 Foursquare announced the possibility of tagging friends belonging to other social networks, i.e., Facebook, even when these are not Foursquare users[5]. In general, there is an increasing concern about the possibility of identifying users from the information that can be extracted from geo-social media.

In this paper, we address the problem of identifying a user through location information from a LBSN. Our aim is to elaborate a number of strategies for the identification of users given their check-in data. More specifically, we firstly propose a trajectory-based approach where a user is identified simply considering the trajectory of spatio-temporal points given by his/her check-in activity. In addition to this, we propose a series of alternative probabilistic Bayesian approaches where a user is characterized by his/her check-in frequency at each location. We also propose to exploit the social ties of the LBSNs by augmenting the frequency information of a user with that of his/her neighbors in the social graph. Finally, we combine the trajectory-based and the frequency-based techniques and propose a hybrid identification strategy. In order to evaluate these techniques, we measure experimentally the loss of victims' privacy as a function of the available anonymized infor-

[1]http://techcrunch.com/2011/12/20/brightkite-winds-down-says-it-will-come-back-with-something-better-again/
[2]http://blog.gowalla.com/
[3]https://www.facebook.com/about/location
[4]https://foursquare.com
[5]http://aboutfoursquare.com/foursquare-extends-friend-tagging-to-facebook/

	l_1	l_2	l_3	l_4
Alice	4	4	4	4
Bob	1	1	1	4
Charlie	5	1	2	0

id	Trace	Other
u_1	l_4, l_1, l_4	s_1
u_2	l_1, l_1, l_1	s_2
u_3	l_1, l_2, l_3	s_3

Table 1: Linking location information across different databases allows the attacker to break users' privacy.

mation. We also propose to quantify the complexity of the identification task by means of the generalized Jensen-Shannon divergence [21] between the frequency histograms of the users.

To the best of our knowledge, this is the first work concerning the problem of identification of users through LBNS location data. We find that the check-in data of the neighbors of a user, depending on the dataset being used, have a limited impact on the ability of identifying that user, which fits with what previous studies have observed on the interaction between mobility and social ties in LBSNs [6, 13, 14]. We also show that the more unique a GPS position is (i.e., the less shared it is among users), the more efficient the trajectory-based strategy is when the number of check-ins that we intend to classify is small. Overall, however, we find that the hybrid approach yields the best classification performance, with an accuracy of more than 90% in some of the selected datasets.

We should stress that the identification strategies proposed in this paper can be generally applied to any setting in which location information and social ties are available. One example is the case of a dataset composed of "significant places" [1] and social connections for a set of users. Significant places of a specific user are usually extracted by means of clustering techniques (see, for example, the seminal work by Ashbrook et al. [1]) and they can be interpreted as his/her check-in locations.

One can argue that by choosing to participate in a LBSN, the user implicitly accepts the respective privacy disclosure agreement. In fact, LBSNs users willingly share their location data on the network, where their identity is publicly visible to all the other users. However, it is possible to note that a potential attacker who intends to break the privacy of an additional source of anonymized location information may use the LBSNs data to transfer the identity information to the anonymized dataset [11]. As a consequence, we believe that it is of pivotal importance to investigate the threats posed by identification attacks of users from their check-in data.

The remainder of this paper is organized as follows. Section 2 defines the identification problem and the motivations for the present work. Section 3 gives an overview of the three datasets selected for this study. In Section 4 we introduce the techniques proposed in this paper for identifying a user given a set of check-ins and we propose a way to measure the complexity of the identification task over a given dataset. In Section 5 we provide an extensive experimental evaluation of the classification accuracy using data from three different LBSNs and we review our main findings and the related work in Section 6. Finally, we conclude the paper in Section 7 and we outline our future research agenda.

2. PROBLEM DEFINITION

We assume that an attacker has access to both unanonymized LBSN data and a source of anonymized location information[6]. This database is anonymized in that the true identities of its participants are replaced by unique random identifiers. Note that such a database may also contain other potentially sensitive data, e.g., health or financial information. Given this setting, the attacker tries to reveal the identities of the participants by linking the location information in the LBSN, where the users' identities are revealed, to the anonymized database.

Let us introduce the problem by means of a toy example illustrated in Table 1. The left part shows, for each user, the number of times that he or she has checked-in at location l_i, whereas the right part shows an additional database of location data in which the identities of the participants have been masked using random identifiers. More specifically, each row of this database consists of an identifier u_i, a sequence of visited locations l_j and an additional sensitive attribute denoted as s_i. The task of the attacker is that of linking the information across the two databases using the location data. In this example, we note that u_1's presence has been recorded 2 out of 3 times at l_4, which suggests that u_1 is either *Alice* or *Bob*, as *Charlie* has never checked-in at l_4. The uncertainty can be further reduced by observing that while the check-in history of *Alice* suggests that she has an equal probability of checking-in at any location, the frequency histogram of *Bob* is sharply peaked at l_4, which fits better the sequence of locations visited by u_1.

Note that the issues that arise from linking information across different databases have been widely investigated in recent years by the community working on differential privacy [33, 26, 11, 7, 22]. The problem we consider in this paper, however, differs from the previous work by being focused on the identity privacy leakage of LBSNs data. With respect to other source of mobility data, in fact, *LBSNs add a further social dimension that can be exploited when trying to break the privacy of an individual*.

3. OVERVIEW OF THE DATASETS

We choose to validate the proposed techniques on three different LBSNs, namely Brightkite, Gowalla and Foursquare. More specifically, we use the Brightkite and Gowalla data collected by Cho et al. [6] and the Foursquare data collected by Gao et al. [13, 14].

The Brightkite data contains 4,491,143 check-ins from 58,228 users over 772,764 location, from April 2008 to October 2010. The Gowalla dataset is composed of 6,442,890 check-ins from 196,591 users over 1,280,969 locations, collected from February 2009 to October 2010. Finally, the Foursquare dataset is a collection of 2,073,740 check-ins from 18,107 users over 43,063 locations, from August 2010 to November 2011. Due to the lack of an API to collect personal check-ins from Foursquare, the authors of [13, 14] collected the data using Twitter's REST API, while the social ties were collected directly from Foursquare. BrightKite and Gowalla instead used to provide an API to directly access the publicly available data.

For each check-in, we have the (anonymized) user identifier, the location identifier, the timestamp and the GPS coordinates where the check-in was made. Note, however, that while in the Foursquare dataset these are precisely the spatial coordinates where the user shared his/her position, in the other datasets these actually refer to the GPS coordinates of the venue itself. As a consequence, the location information in the Foursquare dataset is in a sense much more *unique* [9] than in the other two datasets. By uniqueness, we mean the extent to which a location in a dataset is shared among different individuals, i.e., the less shared a location is, the more unique it is. In this sense, the precise GPS location of a user where he/she performed his/her check-in is more unique than the GPS coordinates of the venue itself, as the latter will be shared in the records of all the users that checked-in at that venue. As a result, the less unique a piece of information is, i.e., the more shared it is among several users, the less discriminative it will be when exploited to identify users.

[6]This could be in the form of check-in data or sequences of GPS points. These can be reduced to a finite set of venues by extracting the set of significant places as in [1].

	SF_B	NY_B	LA_B	SF_G	NY_G	LA_G	SF_F	NY_F	LA_F
number of users	525	494	371	2,203	1,280	690	697	2,592	473
number of check-ins	66,593	61,607	63,923	340,366	136,548	79,616	65,092	258,469	42,011
number of locations	12,929	13,592	11,329	15,673	4,074	2,695	1,173	4,484	1,177

Table 2: Number of users and locations in the the selected cities. The subscript denotes the initial of the name of the LBSN dataset (Brighkite, Gowalla and Foursquare).

Note that, given the nature of our task, identifying users from check-ins scattered all over the world may be considered as an almost trivial task, due to the sparsity of the location information and the lack of a substantial overlap between different users in their check-ins habits. For this reason, we decide to restrict our analysis to the users that are active in San Francisco, New York and Los Angeles, considering only the check-ins in the urban boundaries of these cities. More specifically, given the latitude and longitude of the city center of a city[7], we keep all the users and locations within a 20km radius from it. We select these cities since they have the highest number of active users, so as to render the identification task as hard as possible. Table 2 shows the number of users and locations in each selected city. Note that for each city we consider only the users that performed at least 10 check-ins, as explained in Section 5.

4. IDENTIFICATION METHODS

In this section, we propose a set of techniques to identify a user given a series of check-ins data. Let $C = \{c_1 \dots c_n\}$ denote a set of check-ins. In our dataset, each check-in c_i is labeled with a user identifier u_id_i, a location identifier l_id_i, a timestamp t_i and a GPS point p_i indicating where the user performed the check-in. Let $C(u)$ denote the set of check-ins c_i with $u\_id_i = u$ and $u \in U$, where U is the set of users. For each user u, we divide $C(u)$ into a training test $C_{train}(u)$ and a test set $C_{test}(u)$, where in the latter we remove the user identifier attribute. Given $C_{test}(u)$, our task is that of recovering the identity of the original user. We propose to solve this task by using location data at different levels of granularity. More specifically, we use both the trajectory of high-resolution GPS coordinates visited by the users and the frequency of visits to the different locations. We conclude the section introducing a simple yet effective way to measure the complexity of the identification task over a given dataset.

4.1 Trajectory-based Identification

Since every check-in action is labeled with the precise GPS position where the user was located at that moment, we firstly explore an identification technique based on the analysis of the spatio-temporal information alone. More precisely, let the set of time labeled points p_i in $C_{train}(u)$ and $C_{test}(u)$ be denoted as $T_{train}(u)$ and $T_{test}(u)$ respectively. In other words, $T_{train}(u)$ and $T_{test}(u)$ are spatio-temporal trajectories induced by the check-ins of u. Then, given the spatio-temporal trajectory $T_{test}(u)$, we assign it to the user $v \in U$ who minimizes the distance $dist(T_{train}(v), T_{test}(u))$ defined as follows. Recall that the Hausdorff distance between two finite set of points $A = \{a_1, \cdots, a_m\}$ and $B = \{b_1, \cdots, b_n\}$ is defined as

$$H(A, B) = \max(h(A, B), h(B, A)) \tag{1}$$

where $h(A, B)$ is the directed Hausdorff distance from set A to B

$$h(A, B) = \max_{a \in A} \min_{b \in B} ||a - b|| \tag{2}$$

[7]http://www.census.gov/geo/maps-data/data/gazetteer.html

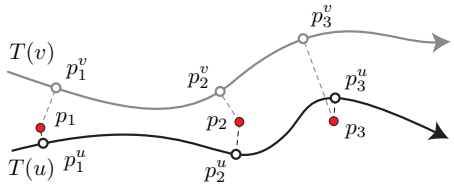


Figure 1: Two users v and u and their traces $T(v)$ (grey) and $T(u)$ (black) along with a set of three points (red) sampled from $T(u)$. These points are classified as belonging to $T(u)$ because the average distance to the corresponding nearest points in $T(u)$ is lower than the average distance to the nearest points in $T(v)$.

and $||\cdot||$ denotes the norm on the underlying space. The modified Hausdorff distance is introduced by Dubuisson et al. [10] as

$$h_m(A, B) = \frac{1}{|A|} \sum_{a \in A} \min_{b \in B} ||a - b|| \tag{3}$$

where $|A|$ denotes the number of points in A. We then define the spatio-temporal distance $d_{st}(p_1, p_2)$ between two points p_1 and p_2 as

$$d_{st}(p_1, p_2) = d_s(p_1, p_2) e^{\frac{d_t(p_1, p_2)}{\tau}} \tag{4}$$

where d_s denotes the distance computed using the Haversine formula [30], while d_t denotes the absolute time difference between two points. Here the exponential is used to smooth the distance between two points according to the absolute difference of their timestamps. Note that by setting $\tau \to \infty$ we ignore the temporal dimension, i.e., the distance between two spatio-temporal points is equivalent to their Haversine distance. As it turns out, due to the spatial and temporal sparsity of the check-in data, the best identification accuracy is achieved for $\tau \to \infty$, and thus we define the distance between a user's trajectory $T_{train}(v)$ and a set of check-in coordinates $T_{test}(u)$ as

$$dist(T_{train}(v), T_{test}(u)) = \frac{1}{|T_{test}(u)|} \sum_{p_1 \in T_{test}(u)} \min_{p_2 \in T_{train}(v)} d_s(p_1, p_2). \tag{5}$$

We stress that the modified Hausdorff distance is not properly a metric, as it is not symmetric. We choose the modified Hausdorff distance over other commonly used distances such as the Hausdorff [32], Fréchet [12], or Dynamic Time Warping distance [3], for its simplicity and robustness to outliers. Note in fact that the Hausdorff distance between $T_{train}(v)$ and $T_{test}(u)$ is low only if every point of either set is close to some point of the other set. This is clearly not true in our case, as we expect $T_{test}(u)$ to contain much fewer points than $T_{train}(v)$, and thus a large portion of $T_{train}(v)$ consists of outliers with respect to $T_{test}(u)$. More specifically, we need to compute the distance between a subset of points and an entire trajectory. The Fréchet and DTW distances, on the other hand, are designed to evaluate the distance between two trajectories of

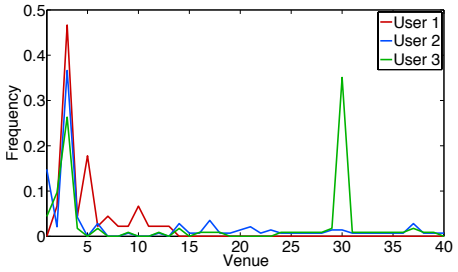


Figure 2: The multinomial models (without Laplace smoothing) for three users in the city of San Francisco.

points, while here the set of points $T_{test}(u)$ can contain as little as a single point. Figure 1 shows the intuition behind the use of the modified Hausdorff distance.

4.2 Frequency-based Identification

Although the GPS points p_i describing a user in the trajectory based model are generally considered to be distinct, they are actually clustered around a limited number of locations. Hence, we can characterize a user with the frequency of visit to this set of locations, rather than the trajectory of spatio-temporal points. In particular, given a set of check-ins $C_{test}(u) = \{c_1 \dots c_m\}$ where the user attribute has been removed, we propose to solve the identification task by selecting the user v which maximizes the posterior probability

$$v^* = \arg\max_{v \in U} P(v|c_1 \dots c_m) \tag{6}$$

where $P(v|c_1 \dots c_m)$ denotes the probability of $v \in U$ being the user who generated the check-in series $C_{test}(u)$.

4.2.1 Multinomial Model

We also develop an identification method based on a multinomial naïve Bayes model, widely used for several classification tasks, such as text classification [25]. By applying Bayes theorem and making the naïve assumption that each check-in c_i is conditionally independent of the others given the user v, we can rewrite Eq. 6 as

$$v^* = \arg\max_{v \in U} P(v) \prod_{i=1}^{m} P(c_i|v) \tag{7}$$

where $P(v)$ is the user prior and $P(c_i|v)$ is the probability of c_i being a check-in generated by v. Here we assume a uniform distribution for the user prior, while we apply a standard maximum likelihood approach to estimate the multinomial distribution associated to each user, i.e.,

$$P(c_i|v) = \frac{N_i^v}{\sum_{j=1}^{n} N_j^v} \tag{8}$$

where N_i^v denotes the number check-ins of v at the location l_id_i in $C_{train}(v)$.

We eliminate zero probabilities by applying Laplace smoothing [24], i.e.,

$$P(c_i|v) = \frac{N_i^v + \alpha}{\sum_{j=1}^{n} N_j^v + \alpha|L|} \tag{9}$$

where $\alpha > 0$ is the smoothing parameter and $|L|$ is the number of locations in our dataset. In other words, we assume a uniform prior over the set of locations. Figure 2 shows the probability distributions over the set of locations of three different users in the city of San Francisco. For the sake of clarity, only the locations visited by at least one of the users are shown.

4.2.2 Time-dependent Multinomial Model

The multinomial model can be enhanced by exploiting the temporal information of the check-ins. In fact, we know that people tend to check-in at the same locations at similar times, yet different people may exhibit different temporal habits. Here, we propose to use 4 time units of 6 hours each to characterise the daily activity of users. Let $\xi \in \Xi = \{1, 2, 3, 4\}$ be a discrete variable denoting the parts of the day. We model each user with 4 different multinomial distributions describing the time dependent check-in frequency over the locations, i.e.,

$$P_\xi(c_i|v) = \frac{N_i^v(\xi) + \alpha}{\sum_{j=1}^{n} N_j^v(\xi) + \alpha|L|} \tag{10}$$

where $P_\xi(c_i|v)$ denotes the time dependent probability of performing a check-in at l_id_i during the time interval ξ and $N_i^v(\xi)$ is the number of check-ins of user v at location l_id_i during the time interval ξ.

4.2.3 Social Smoothing

Given the social nature of LBSNs, it is reasonable to expect that the activity of a user may be influenced by that of his/her friends in the network [14, 13]. Hence, we explore the possibility of exploiting the check-in distributions of the social neighbors of u to augment the previous models. More formally, let $h_u \in \mathcal{R}^n$ be a vector such that $h_u(i)$ denotes the number of check-ins performed by user u at the location i. We first define the similarity between two users u and v as the cosine similarity between h_u and h_v, i.e.,

$$s(u, v) = \frac{h_u^\top h_v}{||h_u|| ||h_v||} \tag{11}$$

where $a^\top b$ denotes the dot product between a and b and $||a||$ is the Euclidean norm of a. The underlying intuition is that the more similar two users are the more likely they are to influence each other.

We then apply a "social smoothing" to the check-in data of v as follows:

$$P(c_i|v) = \frac{N_i^v + \mu \sum_{w \in \mathcal{S}(v)} s(v, w) N_i^w + \alpha}{\sum_{j=1}^{n} N_j^v + \mu \sum_{w \in \mathcal{S}(v)} \sum_{j=1}^{n} s(v, w) N_j^w + \alpha|L|} \tag{12}$$

where $\mathcal{S}(v)$ denotes the social neighborhood of v and μ is a parameter that controls the impact of the social smoothing. The rationale behind the social smoothing is that if a location has not been visited by v, it has a higher chance to be visited in the future if it has been visited by some of his/her friends. However, care should be given to the choice of the value of μ, as large values would introduce too much smoothing, effectively rendering a user indistinguishable from his/her social neighborhood. Note also that we still need to apply Laplace smoothing to avoid zero probabilities.

4.2.4 Hybrid Model

Finally, we propose to merge the spatial and frequency information in a single hybrid model. Given a set of check-ins $C_{test}(u)$ and a user v, we assign the pair a value which is a convex combination of the probability of $C_{test}(u)$ being generated by v with the inverse of the distance to v defined in Eq. 5, i.e.,

$$\gamma(v, C_{test}(u)) = w_{prob} P(C_{test}(u)|v) + \frac{w_{dist}}{1 + dist(T_{train}(v), T_{test}(u))} \tag{13}$$

where w_{prob} and w_{dist} are non negative weights such that $w_{prob} + w_{dist} = 1$. The second term of Equation 13 encodes the spatial similarity between the two trajectories, and it is bounded between 0 and 1. Since we also have that $0 \leq P(C_{test}(u)|v) \leq 1$, it follows that $\gamma(v, C_{test}(u))$ itself will be a real number between 0 and 1.

4.3 Measuring the Complexity of the Identification Task

We conclude this part by introducing a simple yet effective way to quantify the complexity of the identification task over a given dataset, under the assumption that a Bayesian approach is used to break the privacy of the dataset as described in the previous subsection. This in turn requires computing the Jensen-Shannon divergence [21] between the multinomial distributions associated with the users, i.e., their check-in frequency histograms. Unlike other pairwise divergence measures, such as the relative entropy [8], the Jensen-Shannon divergence is designed to deal with $n \geq 2$ probability distributions. Since in our case the number of users n is indeed larger than 2, the choice of the Jensen-Shannon divergence seemed the most appropriate.

Let $P_1, P_2, \cdots, P_n$, with $P_i = \{p_{ij}, j = 1, \cdots, k\}$, be n probability distributions over some finite set X, where $\pi = \{\pi_1, \pi_2, \cdots, \pi_n | \pi_i > 0, \sum \pi_i = 1\}$ is a set of weights, i.e., a set of priors. The generalized Jensen-Shannon divergence of the set $P_1, P_2, \cdots, P_n$ is defined as

$$JS_\pi(P_1, \cdots, P_n) = H(\sum_{i=1}^{n} \pi_i P_i) - \sum_{i=1}^{n} \pi_i H(P_i) \qquad (14)$$

where $H(\cdot)$ denotes the Shannon entropy. Eq. 14 is essentially measuring the irregularity of the set $P_1, \cdots, P_n$ as the difference between the entropy of the convex combination of the P_i and the convex combination of the respective entropies. Interestingly, when all the P_i are equal we have that $JS_\pi = 0$. For the case $n = 2$, Lin [21] has shown that the Jensen-Shannon divergence is bounded between 0 and 1, symmetric and non-negative. However, in the general case where $n > 2$, the upper bound of the Jensen-Shannon divergence becomes $\log(\min(n, k))$ [15].

As a first attempt to rigorously measure the complexity measure of the complexity task, we decided to use the Jensen-Shannon divergence to compute lower and upper bounds of the multiclass Bayes error as shown by Lin [21]. In fact, the Bayes error can be seen as a measure of the hardness of a classification problem. More specifically, the Bayes error estimates the probability of misclassifying an observation in a Bayesian framework, i.e., in our case, the probability of misidentifying an individual. Given a multiclass problem with n classes $c_1, \cdots, c_n$, class conditional distributions $P_1, \cdots, P_n$ and priors $\pi = (\pi_1, \cdots, \pi_n)$, the following relationship between the Jensen-Shannon divergence and the Bayes probability of error $P(e)$ holds:

$$\frac{J_n^2}{4(n-1)} \leq P(e) \leq \frac{J_n}{2}, \qquad (15)$$

where $J_n = H(\pi) - JS_\pi(P_1, \cdots, P_n)$. However, our experimental evaluation found that the bounds to be not tight enough to be informative. In particular, we found the upper bound to be larger than 1, over all the cities and datasets. This may be a consequence of the fact that, in order to reflect the lack of knowledge on the prior probability of the different users, we set $\pi = (\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n})$. Note in fact that $H(\pi) \leq \log(n)$, and in our specific case equality holds. On the other hand, the upper bound of $JS_\pi(P_1, \cdots, P_n)$ is $\log(\min(n, k))$, and, as a consequence, we have that $J_n \leq \log \frac{n}{\min(n,k)}$. In particular, J_n is certainly greater than 1 whenever $k < n$ and $H(\pi) = \log(n)$. Unfortunately, although in our case $n < k$ for all the cities and datasets, we still observe a value of $\frac{J_n}{2} > 1$, thus rendering the bound of limited interest. We also tried to estimate π_i as the frequency of the check-ins of users i with respect to the total number of check-ins, thus lowering $H(\pi)$, but the results were equally uninformative. For example, we found that for the city of New York (Foursquare) $0.011 < P(e) < 2.819$.

Given the limitations of Eq. 15, we propose a different way to measure the complexity of the classification task. Let D be a dataset holding the records of n users, each of which is characterized by a probability distribution P_i over a finite set X of size k. Then the complexity of discriminating the users of D is defined as

$$\mathcal{C}(D) = 1 - \frac{JS_\pi(P_1, \cdots, P_n)}{\log(\min(n, k))}. \qquad (16)$$

Although not directly connected to the Bayes error, $\mathcal{C}(D)$ is bounded between 0 and 1 and it gives us a readily interpretable measure of the complexity of identifying the users of D. More specifically, $\mathcal{C}(D) = 1$ if and only if the values of P_i are equal for all i, i.e., it is impossible to discriminate between the users based on their check-in frequency. Moreover, when the distributions are maximally different, i.e., the frequency vectors P_i form an orthonormal set, then $\mathcal{C}(D) = 0$, i.e., it is trivial to discriminate between the users.

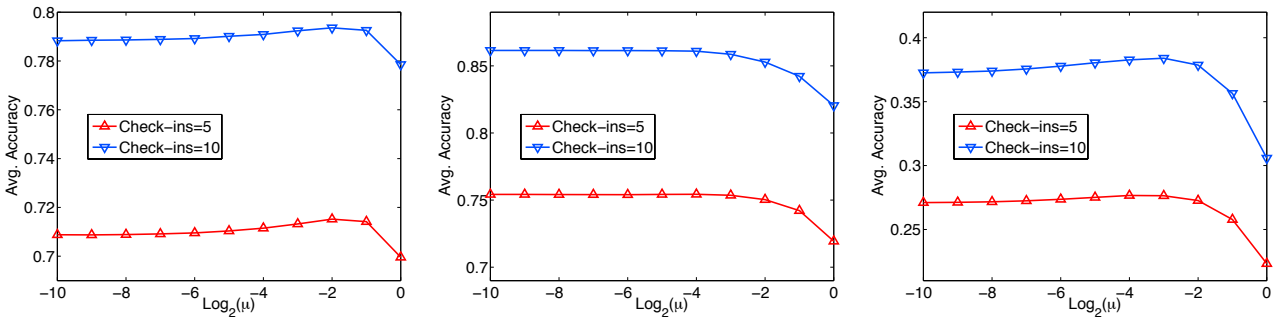
5. EXPERIMENTAL EVALUATION

In this section we will describe the evaluation of the methods presented above. We firstly describe the experimental settings and we then evaluate the performance of the proposed identification strategies.

5.1 Preliminaries

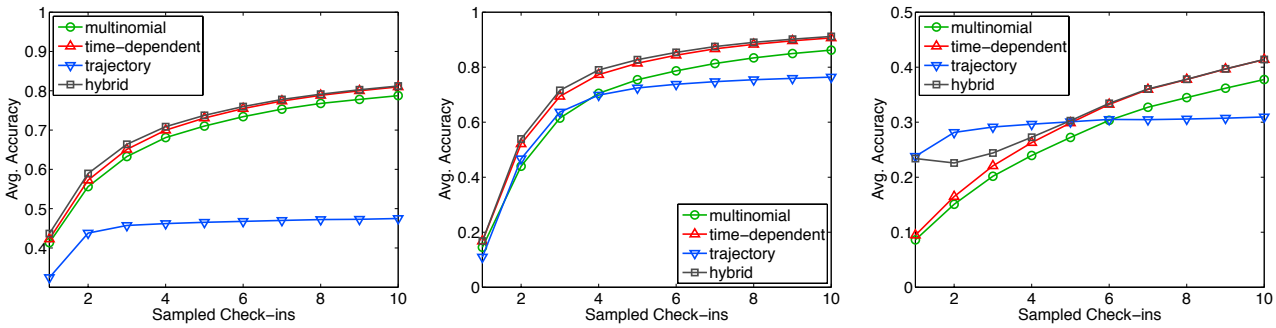
Given a city in our dataset (see Table 2), for each active user we randomly remove 10 check-ins from his/her history $C(u)$ and we use the remaining data to train our algorithms. That is, for each user u we separate $C(u)$ into a training test $C_{train}(u)$ and a test set $C_{test}(u)$. Hence, we are left with $|U|$ sets $C_{test}(u)$ of 10 check-ins, where $|U|$ is the number of users in the city. Given $C_{test}(u)$, the task consists in the identification of the user that originated the set of check-ins. We measure the performance of the different identification strategies in terms of classification accuracy, i.e., the ratio of successfully identified users. Moreover, we are interested in determining the score of each strategy, i.e., the number of guesses required to correctly identify a user. Here the baseline is a random guess, which has average score $|U|/2$. The results of the experiments are then averaged over 100 runs. Note that the scale of the standard error is generally too small to appear in our plots and it has been omitted from the tables as it is always smaller than 10^{-3}. Finally, note that, in the following experiments, we keep the size of $C_{test}(u)$ fixed to 10, but we vary the number of check-ins that we sample from it to identify the users, in order to measure how the performance of the proposed methods depends on the number of observed check-ins. We refer to the set of check-ins sampled from $C_{test}(u)$ as $C_{sample}(u)$.

Recall that the proposed strategies are dependent on the choice of a number of parameters, which include the smoothing parameter α, the social smoothing parameter μ and the interpolation weights w_{prob} and w_{dist}. The parameters are optimized by means of an exhaustive search over a manually defined subset of the parameters space. For each city and dataset, we run our experiments on the training set alone for different combinations of these parameters, and we select the optimal combination in terms of classification accuracy. To this end, we extract 5 check-ins from each user and we



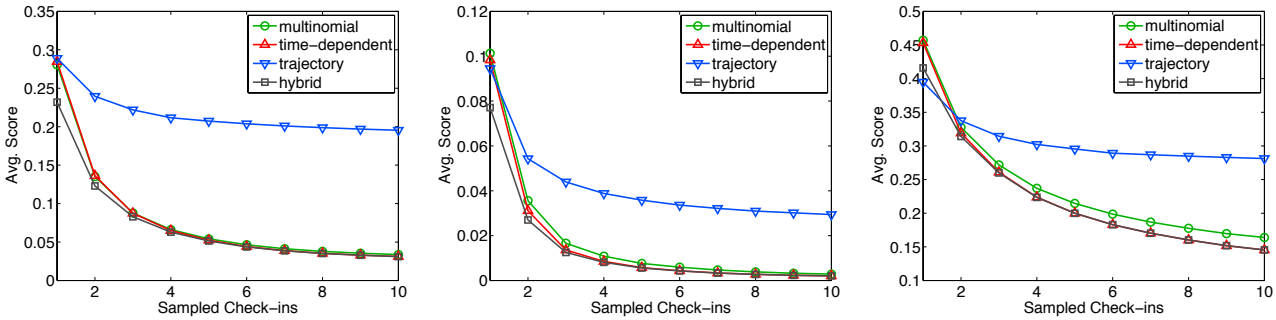
(a) Brightkite (b) Gowalla (c) Foursquare

Figure 3: The effect of the social smoothing on the average classification accuracy for the users in San Francisco.



(a) Brightkite (b) Gowalla (c) Foursquare

Figure 4: The average classification accuracy in the city of San Francisco on the three datasets for increasing size of $C_{sample}(u)$. In the Foursquare dataset, the trajectory-based strategy is the best performing one when the number of sampled check-ins is small. Overall, the hybrid model is the best performing one: it consistently outperforms all the other methods in the Brightkite and Gowalla datasets.



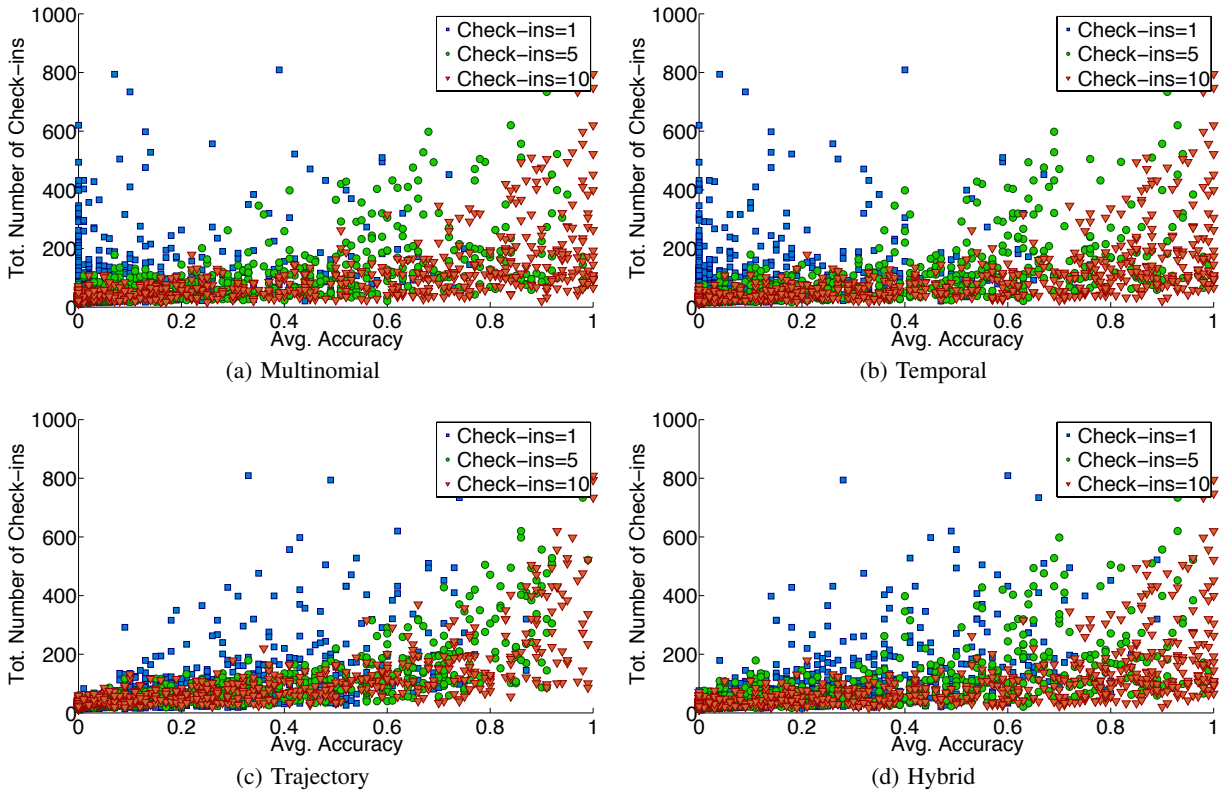
(a) Brightkite (b) Gowalla (c) Foursquare

Figure 5: The average score in the city of San Francisco on the three datasets for increasing size of $C_{sample}(u)$. In terms of average score, the hybrid model consistently outperforms all the other strategies. Also, in the Foursquare dataset the performance gap between the frequency-based strategies and the trajectory-based one is clearly reduced.

apply our identification strategies as described above. Note that, after the test check-ins are removed, the less active users can have as little as 1 check-in in the training set. Thus, we perform the exhaustive search using only those users with more than 5 check-ins in their training set, which in our experimental setting amount for more than 97% of the users. We find that the best classification accuracy is achieved for small values of α. In fact, α represents the prior probability of a user to visit any location in the dataset, independently from his/her check-in history and, therefore, choosing a high value of α would smooth the distribution too much, thus rendering the user harder to classify.

5.2 Experimental Results

Figure 3 shows the effect of applying the social smoothing to the frequency-based strategies. Here we show the average classification accuracy in the city of San Francisco as the value of μ varies. The impact of the social smoothing seems to be rather limited in Foursquare and Brightkite, while in Gowalla the best accuracy is achieved for $\mu = 0$, i.e., when no social smoothing is applied. As



(a) Multinomial (b) Temporal

(c) Trajectory (d) Hybrid

Figure 6: Activity versus average accuracy on the city of San Francisco (Foursquare). Less active users are more difficult to classify correctly, due to the limited number of check-ins available for the training, while very active users are easier to classify.

expected, on the other hand, for large values of μ the performance suddenly drops, as the smoothing starts to render the users indistinguishable from their social neighborhoods. The fact that the social smoothing does not result in a clear increase of the accuracy is not surprising and it fits with what previous studies have observed on the interaction between mobility and social ties in LBSNs [6, 13, 14]. In particular, Cho et al. [6] have found that friendship has a very limited influence on short distance movements (i.e., shorter than 25km, whereas the radius of the cities considered in this paper is 20km), and it is an order of magnitude lower than the influence on long distances (i.e., longer than 1,000km). In particular, they show that only 9.6% of all the check-ins in Gowalla and 4.1% of all the check-ins in Brightkite were first visited by a friend before being visited by a user. In Gao et al. [13, 14], on the other hand, the authors observe an improvement of the location prediction accuracy when the social information is taken into account. However, their study also shows that the impact of the social information is rather limited, and that historical check-in information is more crucial in terms of prediction accuracy.

Figures 4 and 5 show how the average classification accuracy and score on the city of San Francisco vary as we increase the size of $C_{sample}(u)$. The score is reported as a percentage of the baseline score $|U|/2$, i.e., a score of 1 indicates that the method has the same performance of a random guess. We observe that in the Foursquare dataset, when the number of sampled check-ins is smaller than 5, the best performing strategy in terms of accuracy is the trajectory-based one. This is likely due to the high precision and uniqueness of GPS data (the extent to which the data is shared among different users). Recall, in fact, that in this dataset a GPS position refers to the precise spatial coordinates where the user shared his/her position, rather than the coordinates of the venue itself. As a consequence, the spatial information may be sufficient to discriminate among different users who checked-in at the same venue but in positions corresponding to different geographic coordinates, i.e., different places in an urban or non-urban area. However, the same does not hold for Brightkite and Gowalla, where the GPS location of a check-in refers to a unique set of coordinates associated to each venue. In this case, the trajectory-based strategy is always the worst performing one, which confirms our intuition about the uniqueness of the spatial information. As for the frequency-based strategies, we see that the addition of the temporal dimension always yields an increase of the accuracy with respect to the standard multinomial model. Overall, the best performing method is the hybrid one. In the Foursquare dataset, the hybrid model seems to be able to combine the advantages of both the trajectory-based and the frequency-based strategies, by achieving a good performance when the number of sampled check-ins is small, and the best performance when $|C_{sample}(u)| \geq 6$. Conversely, in Brightkite and Gowalla, the hybrid method is consistently outperforming all the others.

In terms of score, Figure 5 also shows that the hybrid method consistently outperforms all the others, in all the three datasets. Note also that, in terms of score, in the Foursquare dataset the advantage of the trajectory-based strategy over the frequency-based one seems to be greatly reduced. In other words, when a user is misclassified, the number of guesses needed to correctly identify him/her is generally higher in the trajectory-based approach than in the frequency-based ones. Interestingly, we also observe that the

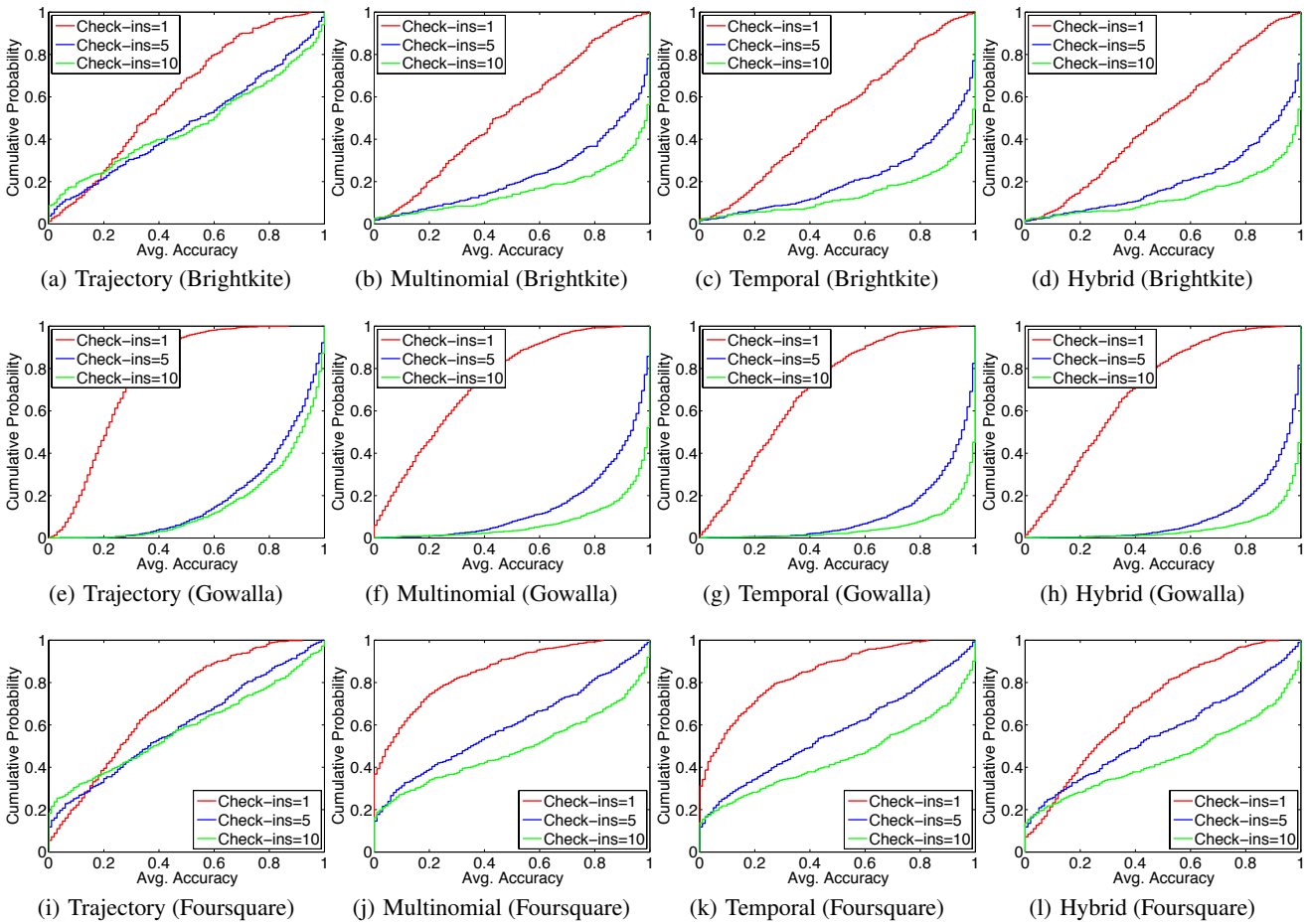


Figure 7: The empirical Cumulative Distribution Function of the user classification accuracies of all the methods on Los Angeles, for the three datasets. In the Gowalla dataset, the hybrid method can identify more than 90% of the users with an accuracy of at least 80%, with $|C_{sample}(u)| = 10$. Note that the results for the other cities show similar trends, and they are omitted due to space constraints.

average scores of the multinomial and its time-dependent version in the other datasets are very close.

The scatter plot of Figure 6 shows the user classification accuracies related to San Francisco (Foursquare), as a function of the users activity. Note that we present the results only for the city of San Francisco due to space limitations. However, the observations we make here hold also in the case of the other cities and datasets. We observe that for the most active users we get a better classification accuracy, as we have a large number of check-ins available to train our models. On the other hand, the performance in terms of classification of less active users can vary considerably. In fact, it would be trivial to identify a user who performed a single check-in at a location where nobody else checked-in. However, a user who performs a small number of check-ins all at very popular venues can be easily misclassified. Figure 6 also shows the advantage of the hybrid method over the other strategies. When the number of sampled check-ins is small, the multinomial and time-dependent models fail to identify most of the users. Instead, the hybrid model shows a distribution similar to that of the trajectory-based strategy, which is the best performing one for small values of $C_{sample}(u)$. When we increase the number of sampled check-ins, on the other hand, the hybrid model performs similarly to the frequency-based strategies, while the trajectory-based approach performs rather poorly, especially for less active users.

Figure 7 shows the empirical distribution function of the user classification accuracies for all the methods and datasets for the city of Los Angeles. These plots show that the classification task seems to be easier on the Brightkite and Gowalla datasets. In the latter, the hybrid method can identify more than 90% of the users with an accuracy of at least 80%, with $|C_{sample}(u)| = 10$. This may be partly due to the fact that, especially in Brightkite, we observe a very large number of locations, which might be the result of fake check-ins. In fact, in both datasets we find several instances of users performing a series of check-ins at locations having different identifiers but same GPS coordinates, and in a relatively short time interval. This in turn results in a very sparse dataset, where there is little overlap between the check-ins of different users, and thus an easier classification task for our strategies. We consider the presence of fake check-ins as a sort of natural feature of datasets extracted from LBSNs. Therefore, we do not perform any preprocessing on our datasets. The classification methodologies have to be robust enough and able to deal with the presence of spurious check-ins associated to a given user.

For the sake of completeness, we report the average classification accuracy and the score of all the strategies over all the cities for the

Trajectory	SF_B	NY_B	LA_B	SF_G	NY_G	LA_G	SF_F	NY_F	LA_F
$\|C_{sample}(u)\| = 1$	0.325	0.402	0.388	0.110	0.189	0.232	**0.238**	0.182	0.299
$\|C_{sample}(u)\| = 5$	0.465	0.493	0.530	0.724	0.730	0.815	0.301	**0.275**	0.402
$\|C_{sample}(u)\| = 10$	0.475	0.505	0.534	0.764	0.760	0.846	0.309	0.294	0.418

Multinomial	SF_B	NY_B	LA_B	SF_G	NY_G	LA_G	SF_F	NY_F	LA_F
$\|C_{sample}(u)\| = 1$	0.413	0.460	0.471	0.145	0.189	0.260	0.086	0.075	0.144
$\|C_{sample}(u)\| = 5$	0.710	0.766	0.769	0.754	0.771	0.850	0.272	0.227	0.404
$\|C_{sample}(u)\| = 10$	0.787	0.837	0.841	0.862	0.867	0.867	0.378	0.301	0.513

Temporal	SF_B	NY_B	LA_B	SF_G	NY_G	LA_G	SF_F	NY_F	LA_F
$\|C_{sample}(u)\| = 1$	0.423	0.467	0.478	0.167	0.224	0.300	0.095	0.079	0.155
$\|C_{sample}(u)\| = 5$	0.731	0.777	0.787	0.814	0.828	0.887	0.299	0.250	0.435
$\|C_{sample}(u)\| = 10$	0.810	0.850	0.860	0.906	0.909	0.948	0.414	0.335	0.552

Hybrid	SF_B	NY_B	LA_B	SF_G	NY_G	LA_G	SF_F	NY_F	LA_F
$\|C_{sample}(u)\| = 1$	**0.437**	**0.487**	**0.496**	**0.168**	**0.225**	**0.301**	0.234	**0.199**	**0.308**
$\|C_{sample}(u)\| = 5$	**0.738**	**0.781**	**0.792**	**0.828**	**0.835**	**0.894**	**0.303**	0.256	**0.439**
$\|C_{sample}(u)\| = 10$	**0.812**	**0.851**	**0.863**	**0.912**	**0.913**	**0.951**	**0.414**	**0.335**	**0.552**

Table 3: Average classification accuracy over all the cities of the three datasets. The best performing method for each city, dataset and size of $C_{sample}(u)$ is highlighted in bold. The standard error is not shown as it was always less than 10^{-3}.

	San Francisco	New York	Los Angeles
Brightkite	0.144	0.079	0.120
Gowalla	0.335	0.279	0.233
Foursquare	0.606	0.571	0.527

Table 4: The identification complexity $\mathcal{C}(D)$ over all the cities and datasets, using the multinomial model.

	San Francisco	New York	Los Angeles
Brightkite	0.083	0.018	0.061
Gowalla	0.190	0.126	0.094
Foursquare	0.469	0.448	0.397

Table 5: The identification complexity $\mathcal{C}(D)$ over all the cities and datasets, using the time-dependent multinomial model.

three datasets in Tables 3 and 6. Again, we see that in most of the cases the best performing method is the hybrid one. Note that we achieve a remarkably high accuracy on some cities: for example, in the city of Los Angeles (Gowalla), we obtain a 95% identification accuracy, when 10 anonymized points are observed. On the other hand, when as little as 1 anonymized point is observed, the maximum accuracy is achieved on the city of Los Angeles (Brightkite), where we can correctly identify nearly 50% of the users.

Finally, we compare these results with those obtained by measuring the identification complexity $\mathcal{C}(D)$ according to Eq. 16. Tables 4 and 5 show the average value of $\mathcal{C}(D)$ for each city and dataset, under the multinomial and time-dependent multinomial models, respectively. More specifically, each time we train the (time-dependent) multinomial model on a training set C_{train} we also compute $\mathcal{C}(C_{train})$. In other words, when computing $\mathcal{C}(C_{train})$ each individual is characterized by a (time-dependent) multinomial distribution p_i, and thus the results should be compared with the classification accuracy of the (time-dependent) multinomial model of Table 3. We should stress, however, that the proposed complexity measure is not restricted to these models and can be applied to any set of probability distributions p_i characterizing an ensemble of users. Finally, note that while in the frequency-based identification methods we applied Laplace smoothing to remove the occurrence of zero probabilities, when computing the Shannon entropy this step is not necessary. In fact, we followed the convention that $0 \log 0 = 0$, which is justified by continuity since $x \log x \to 0$ as $x \to 0$.

Tables 4 and 5 show that the cities in the Foursquare dataset are the most complex ones, which is in accordance with the low classification accuracy achieved by the multinomial model in this dataset. We also observe that the identification task over the cities in the Brighkite dataset seems to be less hard, which is only in partial agreement with the results of Table 3. In fact, when a single point is observed in $C_{sample}(u)$, the Brightkite dataset proves to be the less complex one, in terms of classification accuracy. However, when a larger sample of points is observed, the difference in classification accuracy between Brightkite and Gowalla completely disappears. This may be due to the Laplace smoothing that was applied when training the models. Finally, we observe that the addition of the temporal dimension invariably leads to a reduction of the identification complexity, as already observed in Table 3.

6. DISCUSSION AND RELATED WORK

The results of the experimental evaluation show that it is possible to classify a user from his/her check-in data with high accuracy given a small number of points. In general, the best identification accuracy is achieved by combining frequency and spatial information together. However, if the GPS data refers to the spatial coordinates where the user shared his/her position, the trajectory-based strategy outperforms all the others, when the number of check-ins to classify is small. On the other hand, we observe a negative impact if the GPS information refers to the coordinates of the venue itself, since it has less discriminatory power.

Moreover, in some cases the check-in activity of the friends of a user can be used to increase the identification accuracy, although the effect seems rather limited. The experimental results show that in Brightkite and Gowalla the proposed identification strategies can achieve an accuracy of more than 80% using only 10 check-ins. In Foursquare, we still obtain a classification performance between 30% and 50%, with the same number of check-ins. Given the rising popularity of LBSNs, we believe that our findings raise serious concerns on the privacy of their users. Moreover, we should stress again that the identification strategies proposed in this paper can be

Trajectory	SF_B	NY_B	LA_B	SF_G	NY_G	LA_G	SF_F	NY_F	LA_F
$\|C_{sample}(u)\| = 1$	0.289	0.275	0.222	0.094	0.090	0.063	**0.395**	**0.384**	**0.318**
$\|C_{sample}(u)\| = 5$	0.208	0.180	0.136	0.036	0.031	0.021	0.295	0.271	0.207
$\|C_{sample}(u)\| = 10$	0.195	0.158	0.113	0.029	0.024	0.016	0.281	0.250	0.181

Multinomial	SF_B	NY_B	LA_B	SF_G	NY_G	LA_G	SF_F	NY_F	LA_F
$\|C_{sample}(u)\| = 1$	0.280	0.303	0.250	0.101	0.114	0.085	0.457	0.442	0.376
$\|C_{sample}(u)\| = 5$	0.054	0.044	0.042	0.008	0.009	0.005	0.215	0.230	0.150
$\|C_{sample}(u)\| = 10$	0.034	0.023	0.023	0.003	0.004	0.003	0.164	0.194	0.111

Temporal	SF_B	NY_B	LA_B	SF_G	NY_G	LA_G	SF_F	NY_F	LA_F
$\|C_{sample}(u)\| = 1$	0.285	0.309	0.255	0.098	0.114	0.088	0.453	0.445	0.376
$\|C_{sample}(u)\| = 5$	0.052	0.042	0.040	0.006	0.007	0.004	0.200	0.216	0.141
$\|C_{sample}(u)\| = 10$	0.031	0.022	0.021	0.002	0.003	0.002	0.145	0.179	0.100

Hybrid	SF_B	NY_B	LA_B	SF_G	NY_G	LA_G	SF_F	NY_F	LA_F
$\|C_{sample}(u)\| = 1$	**0.231**	**0.245**	**0.209**	**0.007**	**0.081**	**0.063**	0.415	0.413	0.338
$\|C_{sample}(u)\| = 5$	**0.051**	**0.043**	**0.040**	**0.006**	**0.006**	**0.004**	**0.199**	**0.216**	**0.128**
$\|C_{sample}(u)\| = 10$	**0.031**	**0.022**	**0.021**	**0.002**	**0.002**	**0.002**	**0.145**	**0.178**	**0.099**

Table 6: Average score over all the cities of the three datasets. The best performing method for each city, dataset and size of $C_{sample}(u)$ is highlighted in bold. The standard error is not shown as it was always less than 10^{-3}.

generally applied to any identification problem in which location information and social ties are available. If the location information is in the form of GPS trajectories, it would be sufficient to extract the significant places using clustering techniques [1] and interpret them as the check-in locations.

The advent of mobile technologies has led to several studies concerning human mobility in a geographic space. Recent papers include the prediction of the future location of a person [1], their mode of transport [35] and the identification of individuals from a sample of their location data [9]. In [16] it was shown that there is a high degree of temporal and spatial regularity in human trajectories: users are more likely to visit an area if they have been frequently visited it in the past.

More recently, LBNSs have attracted an increasing interest, due to the massive volume of data generated by their users and their explicit social structure. Examples of applications go from the prediction of the next visited location [27] to the clustering of different types of behaviors of users [18]. Malmi et al. [23] present a transfer learning approach to integrate different types of movement data, including LBSNs check-ins, in order to address the next place prediction problem. Gao et al. [14], on the other hand, propose a geo-social correlation model to capture check-ins correlations between users at different geographical and social distances. Interestingly, they find that there is a higher correlation between users who are not friends but live in the same area rather than direct friends. Similarly, Cho et al. [6] study how the friendships in LBSNs can influence human mobility, and find that in general the influence is higher on long-range movements rather than short-range ones. In another paper, Gao et al. [13] propose a series of models that integrate social information in a location prediction task. Joseph et al. propose to use Latent Dirichlet Allocation to model the check-in activity of Foursquare users and cluster them into different groups with different interests [18]. Vasconcelos et al. [34] investigate the use of "tips", "dones" and "todos" in Foursquare to cluster users profiles. The problem of privacy in LBSNs is discussed and analyzed in Ruiz et al. [31]. In particular, the authors study a number of privacy issues related to the location and identity of LBSN users, and describe possible means of protecting privacy.

Finally, Pontes et al. [29, 28] focus on the inference of the user home location using publicly available information from Foursquare and two different online social networks, namely Google+ and Twitter. More specifically, in [29] the authors show that it is possible to infer with high accuracy where a user lives based on his or her set of Foursquare activities (such as "todos"). In [28], the analysis is extended to Google+ and Twitter, where a number of attributes including the location of the users' friends are used to infer the home city as well as their residence location of the individuals.

With respect to this body of work, to the best of our knowledge, our paper is the first attempt of studying the problem of user identification from LBSNs data. We believe that this issue will be increasingly important, given the ever growing popularity of smartphones running a plethora of location-aware (and usually socially-aware) applications.

7. CONCLUSIONS AND FUTURE WORK

In this paper we have introduced and evaluated a series of techniques for the identification of users in LBSNs. We have tested the proposed strategies using three datasets from different LBSNs, namely Brightkite, Gowalla and Foursquare. We have showed that both the GPS information contained in a user's check-ins and the frequency of visits to certain locations can be used to successfully identify him/her. In particular, we have demonstrated that it is possible to achieve a high level of accuracy with only 10 check-ins, thus raising serious concerns with respect to the privacy of LBSNs users. Finally, we have proposed a simple yet effective way to quantify the complexity of the identification task over a given dataset.

We plan to apply the proposed methods on different datasets, since we are aware of the possible peculiarities and limitations of those used in this study. Indeed, even if we believe that the proposed methodology can be applied to a vast number of identification problems for which geographic and social information are available, we aim to investigate the generalizability of the identification strategies presented in this work to larger and more challenging datasets, which may thus demand more scalable and efficient machine learning techniques. Our future research agenda also includes the definition, implementation and evaluation of obfuscation techniques based on the findings presented in this paper. We also intend to investigate the use of our identification complexity measure on different datasets and to extend it to more general sce-

narios, considering also additional information from users' profiles, if available.

Acknowledgement

This work was supported through the EPSRC Grant "The Uncertainty of Identity: Linking Spatiotemporal Information Between Virtual and Real Worlds" (EP/J005266/1)

8. REFERENCES

[1] D. Ashbrook and T. Starner. Using GPS to Learn Significant Locations and Predict Movement Across Multiple Users. *Personal and Ubiquitous Computing*, 7(5):275–286, 2003.

[2] A. R. Beresford and F. Stajano. Location privacy in pervasive computing. *IEEE Pervasive Computing*, 2(1):46–55, 2003.

[3] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *Proceedings of the AAAI-94 Workshop on Knowledge Discovery in Databases*, volume 10, pages 359–370. Seattle, WA, 1994.

[4] C. Bettini, X. S. Wang, and S. Jajodia. Protecting privacy against location-based personal identification. In *Secure Data Management*, pages 185–199. Springer, 2005.

[5] J. Bohn, V. Coroamă, M. Langheinrich, F. Mattern, and M. Rohs. Social, economic, and ethical implications of ambient intelligence and ubiquitous computing. In *Ambient Intelligence*, pages 5–29. Springer, 2005.

[6] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: User movement in location-based social networks. In *Proceedings of SIGKDD'11*, pages 1082–1090. ACM, 2011.

[7] C.-Y. Chow and M. F. Mokbel. Trajectory privacy in location-based services and data publication. *ACM SIGKDD Explorations Newsletter*, 13(1):19–29, 2011.

[8] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[9] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, 3, 2013.

[10] M.-P. Dubuisson and A. K. Jain. A Modified Hausdorff Distance for Object Matching. In *Proceedings of ICPR'94*, pages 566–568, 1994.

[11] C. Dwork. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation*, pages 1–19. Springer, 2008.

[12] T. Eiter and H. Mannila. Computing Discrete Fréchet Distance. Technical report, Technische Universitat Wien, 1994.

[13] H. Gao, J. Tang, and H. Liu. Exploring social-historical ties on location-based social networks. In *Proceedings of ICWSM'12*, 2012.

[14] H. Gao, J. Tang, and H. Liu. gSCorr: modeling geo-social correlations for new check-ins on location-based social networks. In *Proceedings of CIKM'12*, pages 1582–1586. ACM, 2012.

[15] J. F. Gómez-Lopera, J. Martínez-Aroza, A. M. Robles-Pérez, and R. Román-Roldán. An analysis of edge detection by using the jensen-shannon divergence. *Journal of Mathematical Imaging and Vision*, 13(1):35–56, 2000.

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

[17] M. Gruteser and D. Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In *Proceedings of MobiSys'03*, pages 31–42. ACM, 2003.

[18] K. Joseph, C. H. Tan, and K. M. Carley. Beyond Local, Categories and Friends: Clustering Foursquare Users with Latent Topics. In *Proceedings of UbiComp'12*, pages 919–926. ACM, 2012.

[19] P. Kalnis, G. Ghinita, K. Mouratidis, and D. Papadias. Preventing location-based identity inference in anonymous spatial queries. *IEEE Transactions on Knowledge and Data Engineering*, 19(12):1719–1733, 2007.

[20] J. Krumm. A survey of computational location privacy. *Personal and Ubiquitous Computing*, 13(6):391–399, 2009.

[21] J. Lin. Divergence Measures based on the Shannon Entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[22] C. Y. T. Ma, D. K. Y. Yau, N. K. Yip, and N. S. Rao. Privacy vulnerability of published anonymous mobility traces. *IEEE/ACM Transactions on Networking*, 21(3):720–733, 2013.

[23] E. Malmi, T. M. T. Do, and D. Gatica-Perez. From Foursquare to My Square: Learning Check-in Behavior from Multiple Sources. In *Proceedings of ICWSM'13*, 2013.

[24] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[25] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In *Proceeding of the AAAI-98 Workshop on Learning for Text Categorization*, volume 752, pages 41–48, 1998.

[26] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *Proceedings of SP'08*, pages 111–125. IEEE, 2008.

[27] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo. A tale of many cities: Universal patterns in human urban mobility. *PLOS ONE*, 7(5):e37027, 2012.

[28] T. Pontes, G. Magno, M. Vasconcelos, A. Gupta, J. Almeida, P. Kumaraguru, and V. Almeida. Beware of what you share: Inferring home location in social networks. In *Proceedings of ICDM'12 Workshops*, pages 571–578. IEEE, 2012.

[29] T. Pontes, M. Vasconcelos, J. Almeida, P. Kumaraguru, and V. Almeida. We Know Where you Live: Privacy Characterization of Foursquare Behavior. In *Proceedings of UbiComp'12*, pages 898–905. ACM, 2012.

[30] C. C. Robusto. The Cosine-Haversine formula. *The American Mathematical Monthly*, 64(1):38–40, 1957.

[31] C. Ruiz Vicente, D. Freni, C. Bettini, and C. S. Jensen. Location-related privacy in geo-social networks. *IEEE Internet Computing*, 15(3):20–27, 2011.

[32] J.-R. Sack and J. Urrutia. *Handbook of Computational Geometry*. North Holland, 1999.

[33] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[34] M. A. Vasconcelos, S. Ricci, J. Almeida, F. Benevenuto, and V. Almeida. Tips, Dones and Todos: Uncovering User Profiles in Foursquare. In *Proceedings of WSDM'12*, pages 653–662. ACM, 2012.

[35] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma. Understanding Mobility based on GPS Data. In *Proceedings of UbiComp'08*, pages 312–321. ACM, 2008.