Mobility Profiling for User Verification with Anonymized Location Data

Miao Lin[1], Hong Cao[2], Vincent Zheng[3], Kevin Chen-Chuan Chang[3], Shonali Krishnaswamy[1]
[1]Institute for Infocomm Research, A*STAR, Singapore
[2]McLaren Applied Technologies, APAC
[3]Advanced Digital Sciences Center, UIUC, Singapore
[1]{linm, spkrishna}@i2r.a-star.edu.sg, [2]hong.cao@mclaren.com,
[3]vincent.zheng@adsc.com.sg, [3]kcchang@illinois.edu

Abstract

Mobile user verification is to authenticate whether a given user is the legitimate user of a smartphone device. Unlike the current methods that commonly require users active cooperation, such as entering a short pin or a one-stroke draw pattern, we propose a new passive verification method that requires minimal imposition of users through modelling users subtle mobility patterns. Specifically, our method computes the statistical ambience features on WiFi and cell tower data from location anonymized data sets and then we customize Hidden Markov Model (HMM) to capture the spatial-temporal patterns of each user's mobility behaviors. Our learned model is subsequently validated and applied to verify a test user in a time-evolving manner through sequential likelihood test. Experimentally, our method achieves 72% verification accuracy with less than a day's data and a detection rate of 94% of illegitimate users with only 2 hours of selected data. As the first verification method that models users' mobility pattern on location-anonymized smartphone data, our achieved result is significant showing the good possibility of leveraging such information for live user authentication.

1 Introduction

Smartphones nowadays have become important sensing and personal assistants to support a wide spectrum of user daily activities from communication, scheduling, social networking, reading, shopping to entertainment. People carry their smartphone wherever they go and constantly interact their devices. As a smartphone records rich private information of one's activities, their security cannot be taken for granted. The existing user verification methods, such as entering a short personal identification number (PIN), drawing a one-stroke pattern [von Zezschwitz *et al.*, 2013], require users' active cooperation to set up and to memorize a numeric or graphic draw pattern, and enter it for authentication whenever it is required. Even though such methods have been widely adopted, it is worth noting that the imposition on users to memorize and manage their secret credential, and enter it all the times is a significant burden. As a result, many choose not to use any on-access verification option. Even for those who use PIN and draw pattern, such information can be easily shoulder surfed, leading to the loss of security.

In order to complement existing user verification methods mentioned earlier, we explore in this paper a new way of user verification, i.e., through passively modelling and monitoring the subtle behavioral patterns (mobility in our case) of the users on the ubiquitous location-anonymized data. Our choice of characterizing mobility patterns is motivated by the fact, as pointed out in [de Montjoye *et al.*, 2013], that individuals present distinctive mobility patterns in terms of the spatial-temporal transitions and the daily route. Also, a smartphone is known to continuously receive various location-relevant information such as cell towers [Song *et al.*, 2010b; 2010a] and WiFi access points [Zheng *et al.*, 2008] to support the basic communication needs. This provides a basis for passive and continuous user verification through profile modelling on their mobility patterns.

Previously, relevant research works have been carried out in the areas of mobile user profiling and user identification. **For mobile user profiling**, the works in [De Mulder *et al.*, 2008; Bayir *et al.*, 2009] profiled individuals' mobility patterns using spatial-temporal data. Mulder *et al.* [De Mulder *et al.*, 2008] modeled individuals' mobility by the first-order Markov model based on cell tower data, and each user's mobility pattern was represented in terms of the transition between locations and their stationary distribution. The profile model was applied in the identification setting to show that the individuals can be simply deanonymized based on the historical sequence of GSM towers that the user has visited. Bayir *et al.* [Bayir *et al.*, 2009] learned individuals' mobility profile using the frequent path information between different cell towers where the frequent patterns of the mobility profiles were searched by the AprioriAll algorithm [Agrawal and Srikant, 1995]. The profile model was used for analyzing crowd mobility patterns, e.g., for the sake of city planning. **For mobile user verification/identification**, a few studies [de Montjoye *et al.*, 2013; Gambs *et al.*, 2013; Lin *et al.*, 2015] leveraged on either the geographic data or the smartphone sensory data. Montjoye *et al.* [de Montjoye *et al.*, 2013] aimed to identify individuals based on their spatial-temporal points. They found that 95% of mobile users in a population of 1.5 million could be correctly identified by using only four spatio-temporal points extracted for each

user. Gambs *et al.* [Gambs *et al.*, 2013] constructed individuals' mobility models based on the Mobility Markov Chain (MMC), and further applied the MMC in linking users' traces in the training phase and the testing phase. Based on WiFi and cell tower data, Lin *et al.* [Lin *et al.*, 2015] proposed to use both dynamical time warping and k-nearest neighbor classifier to solve user identification/verification as multi-class classification for fixed-duration data.

Most of the studies [de Montjoye *et al.*, 2013; Gambs *et al.*, 2013] compared users' mobility profiles based on the explicit location symbols, which is not applicable in the anonymized data set. This is because in the anonymized data set each user's data is done separately, making two user's data unlinkable using explicit symbols. Also, none of the above methods [De Mulder *et al.*, 2008; Bayir *et al.*, 2009; de Montjoye *et al.*, 2013; Gambs *et al.*, 2013; Lin *et al.*, 2015] was designed from the pure perspective of user verification. Thus, they cannot be applied in our scenario of user verification using anonymized location data.

Our methodology takes the following three parts. First, we describe users' mobility using statistical ambience features of the ubiquitous cell tower and WiFi data. Our method is applied to the location anonymized data set that does not contain any explicit location information. As the explicit location information are considered highly private, it is common that the locations and its relevant substitutes are dynamically anonymized into a discrete set of symbols before sharing the data set. As we do not directly use the explicit location information in our modelling, it is safer to share our model in the cloud for centralized online verification service. Second, we profile users' mobility by using HMM based on the extracted statistical features. The reasons of using HMM are given as follows. 1) Individuals' mobility data are sequential data and present strong temporal relevance. Generally, we can describe each users' mobility in a repeatable regular basis, e.g., the transition from home to working place, the periods of staying at home, etc. HMM is suitable to model the mobility in such a case. 2) In the verification phase, the testing data also come in sequentially, thus the verification method should output current result by accumulating the results from previous data. HMM is suitable for the case of reusing the previous results, e.g., calculating the probability of the sequence, according to forward-backward algorithm (the details are given in Section 3.2). Third, our user verification is performed in a time-evolving manner through sequential probability ratio test.

Figure 1 shows the framework of our experimental methodology in three phases, training, validation and verification. **In the Training phase,** we tailor HMM to learn each user's mobility profile based on the time series of user implicit mobility features and statistical ambience features. The implicit mobility features include the temporal information and the very existence of each type of observations. For instance, users may turn off WiFi scanner to save the power when they are in transitions. Thus, the missing of WiFi information can implicitly indicate user's mobility states. The statistical ambient features are used to describe the network circumstances that individuals stay in. **In the validation phase,** we construct each user's validation model, including one genuine model and a list of background models. The genuine model is used to characterize the owner user through the trained HMM from the training phase. The list of the background models are used to model the data from different imposters, including those presenting either similar mobility patterns or dissimilar mobility patterns to the legitimate user [Sahoo *et al.*, 2012]. **In the verification phase,** we verify whether the given phone user is the legitimate owner or not by using sequential probability ratio test based on 1) a sequence from an unknown user, 2) the profile and validation model from the claimed user.

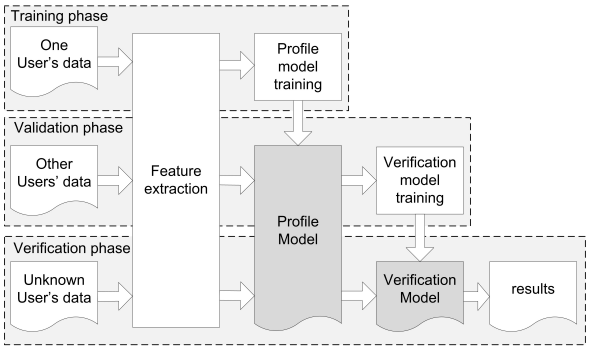


Figure 1: Overview of the user verification framework.

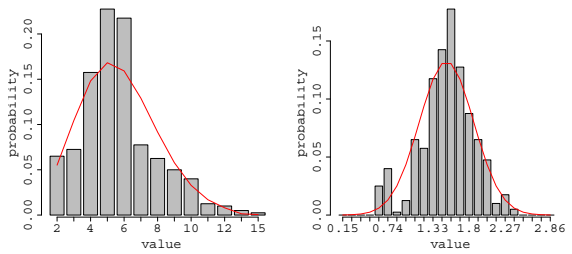
Our contributions of this study are three fold. First, we identify a new problem in terms of 1) passively verifying smartphone users and 2) using the anonymized location symbols. The proposed method can complement existing smartphone protection methods by providing a more secure user verification mechanism. Second, because the anonymized location symbols are inconsistent among different users' data, we propose to use statistical ambience features to describe users' mobility behaviors, and this method can work well on the common location-anonymized data in the cloud or server. Third, we justify our choice of using the HMM to learn individuals' profiles using two types of the location data. Last, we evaluate our mobile user verification method by using two data sets extracted from the device analyzer data set. Experimental results show that our method achieves 72% accuracy with less than a day's data and a detection rate of 94% of illegitimate users with only 2 hours of selected data.

2 Mobile User Profiling

In this section, we present the methodology of how to tailor Hidden Markov Model to learn each user's mobility profile based on anonymized location data.

Before describing our model in details, we introduce some notations. Let $\boldsymbol{O}_t = \{H_t, E_t, \boldsymbol{O}_t^c, \boldsymbol{O}_t^a, \boldsymbol{O}_t^{ac}\}$ denote the observation at the t^{th} hour, where H_t denotes the temporal information including hour of the day and day of the week, E_t denotes the existence of each type of the observations, $\boldsymbol{O}_t^c$ denotes the detailed cell tower observation, $\boldsymbol{O}_t^a$ denotes the detailed WiFi observation, and $\boldsymbol{O}_t^{ac}$ denotes the WiFi connection information.

Specifically, in the t^{th} hour, the cell tower observation is given as $\boldsymbol{O}_t^c = \{n_{t,i}\}$, where each $n_{t,i}$ denotes the number of connections to the i^{th} cell tower observed in this hour. The



(a) (b)

Figure 2: (a) The distributions of number of distinct cell towers in a given state fitted by a Poisson distribution. (b) The distribution of the entropy of the cell towers in a given state fitted by a Gaussian distribution.

WiFi observation is given as $\boldsymbol{O}_t^a = \{m_t, \{m_{t,i}\}\}$, where m_t denotes the total number of scans during this hour and $\{m_{t,i}\}$ denotes the number of scans for each of the WiFi APs in this hour. Note that, not every WiFi AP is shown in each scanned result, thus $0 \leq m_{t,i} \leq m_t$. If the smartphone connects to WiFi AP(s), $\boldsymbol{O}_t^{ac}$ denotes the symbol(s) of these WiFi AP(s), otherwise it is empty.

In the Hidden Markov model, let $\Omega = \{1, 2, .., K\}$ denote a list of one user's mobility states, and the mobility state at the t^{th} hour as S_t, where $S_t \in \Omega$. In the following, we detail the modeling of users' mobility.

Implicit mobility features. There are four implicit user mobility features, namely, temporal information, the existence of the cell tower observation, the WiFi observation, the connection to a WiFi AP.

Let H_t denote the temporal information including hour of the day h_t and day of the week w_t, where $h_t \in \{0, 1, 2, ..., 23\}$ and $w_t \in \{0, 1, 2, ..., 6\}$. Since individuals may not have regular mobility behaviors exactly in the same hour, we divide the hour space into a few non-overlapping space of equal size. The hour information is modeled by a probability mass function over the space. Similarly, we apply the same method for the temporal information day of the week.

Let $E_t = \{E_t^c, E_t^a, E_t^{ac}\}$ denote the existence of each observation, where $E_t^c \in \{0, 1\}$ corresponds to the existence of the cell tower observation, $E_t^a \in \{0, 1\}$ corresponds to the WiFi observation, and $E_t^{ac} \in \{0, 1\}$ corresponds the connection to a WiFi AP or not. Let $\varepsilon_k = \{\varepsilon_{k,c}, \varepsilon_{k,a}, \varepsilon_{k,ac}\}$ denote the parameter set of describing the existence features given state k. Thus,

$$P(E_t^c = 1 | S_t = k) = \varepsilon_{k,c} \tag{1}$$

where $\boldsymbol{O}_t^c \neq \emptyset$, $0 \leq \varepsilon_{k,c} \leq 1$, and it is similar for the other two existence features.

Statistical ambience features. There are two types of statistical ambient features extracted from either cell tower observation and WiFi observation, namely, distinct number of items, and the entropy of these items.

By grouping the cell tower data into different states, we generate the distribution of the cell tower features. Figure 2 shows the number of distinct cell towers is well-fitted by a Poisson distribution, and the entropy of these cell towers is well-fitted by a Gaussian distribution. In the following, we model these two types of features by Poisson distributions and Gaussian distributions, respectively. Specifically, in the cell tower observation, the first feature, given as the number of distinct cell tower, is $F_t^{(1)} = n_{t,d} = |\{n_{t,i}\}|$, and it is modeled by a Poisson distribution. The second feature, given as the entropy of the cell towers, is $F_t^{(2)} = n_{t,e}$, where $n_{t,e} = \sum_i \hat{p}_{t,i} \log \hat{p}_{t,i}$ and $\hat{p}_{t,i} = \frac{n_{t,i}}{\sum_i n_{t,i}}$, and it is modeled by a Gaussian distribution. The probability of the cell tower observation given state k in the t^{th} hour is

$$P(\boldsymbol{O}_t^c | S_t = k) = \prod_{i=1}^{2} P(F_t^{(i)} | C_t = 1, S_t = k) \tag{2}$$

where $C_t = 1$ denotes current observation is cell tower.

Similarly in the WiFi observation, the total number of distinct WiFi APs observed is $F_t^{(1)} = m_{t,d} = |\{m_{t,i}\}|$, and it is modeled by a Poisson distribution. The entropy of these WiFi APs is $F_t^{(2)} = m_{t,e} = \sum_i \hat{p}_{t,i} \log \hat{p}_{t,i}$, where $\hat{p}_{t,i} = \frac{m_{t,i}}{m_t}$, and it is modeled by a Gaussian distribution. Similarly, the probability of the WiFi observation given state k in the t^{th} hour follows Eq. (2) given $C_t = 0$.

In summary, the parameter sets in the model are $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{R}, \boldsymbol{\theta}^m, \boldsymbol{\theta}^c, \boldsymbol{\theta}^a\}$, where $\boldsymbol{\pi}$ and $\boldsymbol{R}$ are the parameters describing the initial probability and transition matrix of the states, $\boldsymbol{\theta}^m$ denotes the parameter sets of describing the implicit user mobility features, $\boldsymbol{\theta}^c$ and $\boldsymbol{\theta}^a$ denotes the parameter sets of describing the statistical ambient item features of cell tower and WiFi APs, respectively.

Modeling users' mobility. In the proposed HMM, the observation probability is

$$\begin{aligned} P(\boldsymbol{O}_t | S_t) &= P(H_t, E_t, \boldsymbol{O}_t^c, \boldsymbol{O}_t^a, \boldsymbol{O}_t^{ac} | S_t) \\ &= P(H_t | S_t) P(E_t | S_t) P(\boldsymbol{O}_t^c | S_t) P(\boldsymbol{O}_t^a | S_t) \end{aligned}$$

where the connection observation $\boldsymbol{O}_t^{ac}$ is modeled in the existence feature E_t^{ac}. Except the difference of the observation probability, the other parts, e.g., state transitions, initial probability, are the same as the traditional HMM.

With a given number of mobility states K, the inference of the parameter set $\boldsymbol{\theta}$ is performed via the EM algorithm, similar to [Dempster *et al.*, 1977]. Also, the proposed method is extendable in the case of using both types of the observation or either type of the observation. Specifically, the method described in this section provides a framework of using both types of the observations, and the learned Hidden Markov Model is denoted by B-HMM. We can train the proposed Hidden Markov model only using one type of the observation, namely, C-HMM trained using cell tower observation and A-HMM trained using WiFi observation. In the experimental section, we compare the performance of using different HMM in user verification.

3 User Verification

In this section, we describe both the validation and verification shown in Figure 1.

3.1 Validation

In the validation phase, for each user we construct the validation model using one genuine model and a list of background models. The genuine model is the trained hidden Markov model for the smartphone owner. The list of background models are the other users' trained hidden Markov models, including a list of the most similar ones to the current genuine model and the same number of the most dissimilar ones. The similar ones help to differentiate the traces from users with similar mobility behaviors. The dissimilar ones help to identify the traces from those with very different mobility behaviors. This is because if all the background models are very similar to the user's genuine model, the mobility traces from those with very different patterns will be badly fitted by either genuine model or any of the background models, leading to an unreliable verification result [Bimbot *et al.*, 2004].

In order to get a list of similar models and dissimilar models to the current profile model, we compute the distance between HMMs according to the method given in [Do, 2003]. The distance between two HMMs $\mathcal{M} = \{\boldsymbol{\pi}, \boldsymbol{R}, \boldsymbol{\theta}^m, \boldsymbol{\theta}^c, \boldsymbol{\theta}^a\}$ and $\tilde{\mathcal{M}} = \{\tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{R}}, \tilde{\boldsymbol{\theta}}^m, \tilde{\boldsymbol{\theta}}^c, \tilde{\boldsymbol{\theta}}^a\}$ is bounded by $D(\mathcal{M}||\tilde{\mathcal{M}}) \leq \sum_{j=1}^{K} v_j(D(\boldsymbol{r}_j||\tilde{\boldsymbol{r}}_j) + D(\boldsymbol{b}_j||\tilde{\boldsymbol{b}}_j))$, where v_j is the stationary probability of state j. $D(\boldsymbol{r}_j||\tilde{\boldsymbol{r}}_j)$ is the KL distance [Kullback and Leibler, 1951] between the transition vector given state j. $D(\boldsymbol{b}_j||\tilde{\boldsymbol{b}}_j)$ is the KL distance between the corresponding observation distribution given state j. And we use the upper bound to approximate the distance of two HMMs.

3.2 Verification

In this part, we show how to conduct sequential probability ratio test (S-test) [Wald, 1945] in user verification. The sequential probability ratio test is specifically designed for testing the sequence with increasing length and this test makes the decision in a time-evolving manner by accumulating the results from previous step.

Let $\boldsymbol{O}_{1:T}$ be the mobility sequence that needs to be verified whether it belongs to user u or not. For user u, we have the profile model $\mathcal{M}_0$ and the background model lists $\{\mathcal{M}_1\}$. We would like to verify the following two hypotheses:

- H_0 : the sequence $\boldsymbol{O}_{1:T}$ is from the claimed user u.
- H_1 : the sequence $\boldsymbol{O}_{1:T}$ is not from the claimed user u.

Sequential probability ratio test (S-test). There are two thresholds in this test, namely, ϵ_l and ϵ_h, where $\epsilon_l < 0 < \epsilon_h$. S-test consists a list of tests based on subsequences. The first test starts from Index 1 and ends at Index i, where in our case $i \geq 2$ since we need at least two temporally-adjacent samples to capture the dependency of the observations. Then, we calculate the likelihood ratio given current subsequence $\boldsymbol{O}_{1:i}$,

$$\Lambda(\boldsymbol{O}_{1:i}) = \log P(\boldsymbol{O}_{1:i}|\{\mathcal{M}_1\}) - \log P(\boldsymbol{O}_{1:i}|\mathcal{M}_0) \quad (3)$$

The decision is made as follows

$$\begin{cases} Accept \quad H_0, & \text{If } \Lambda(\boldsymbol{O}_{1:i}) < \epsilon_l \\ Accept \quad H_1, & \text{If } \Lambda(\boldsymbol{O}_{1:i}) > \epsilon_h \\ Undetermined, & Otherwise \end{cases} \quad (4)$$

If these is no decision given current subsequence, we increase the length of the subsequence by one and conduct the test again. We terminate the test either there is a decision or we reach the end of the sequence. The theory of S-test provides the criteria of choosing ϵ_l and ϵ_h according to the predefined missed detection rate P_{misD} and false alarm rate P_{FA} [Wald, 1945], where $\epsilon_l \approx \log \frac{P_{misD}}{1-P_{FA}}$ and $\epsilon_h \approx \log \frac{1-P_{misD}}{P_{FA}}$. In the experiments, we set these two thresholds in S-test ϵ_l and ϵ_h as the default values according to $P_{misD} = 0.01$ and $P_{FA} = 0.01$. This is because assigning the relatively low missed detection rate and false alarm rate will impose high constraint on the decision of the tests, thus it can lead to more accurate results.

Algorithm 1 User verification algorithm.

Input: 1) a mobility sequence $\boldsymbol{O}_{1:T}$ from an unknown user, 2) a claimed user ID, 3) two threshold ϵ_l and ϵ_h. **Output:** verification result.

```
Initialize t = 2;
while t <= T & flag == Un do
    Initialize flag = Un, p_gen = 0, p_bac = 0;
    for j ∈ {0, 1} do
        Get current model(s) M_j and model number num;
        for m ∈ {1, num} do
            if t == 2 then
                Calculate {α_k(t); t ∈ [1, 2], k ∈ [1, K]} according to Eq.(5) and Eq.(6);
            else
                Update the forward probability matrix by adding {α_k(t); k ∈ [1, K]} according to Eq.(6);
            end if
            Calculate P(O_{1:t}) = Σ_k α_k(t);
            if j == 0 then
                p_gen = P(O_{1:t});
            else
                p_bac = max{p_bac, P(O_{1:t})};
            end if
        end for
    end for
    Do verification using p_gen and p_bac according to Eq.(3) and Eq. (4);
end while
return flag
```

Verification method. Due to the nature of the hidden Markov Model, we can re-use the test results based on previous subsequence in S-test. Specifically, let $\alpha_k(t)$ be the forward probability of observing the first t samples given state k at time t, thus $\alpha_k(t) = P(\boldsymbol{O}_{1:t}, S_t = k)$ and it can be calculated recursively,

$$\alpha_k(1) = \pi_k P(\boldsymbol{O}_1|S_1 = k) \quad (5)$$

$$\alpha_k(t) = P(\boldsymbol{O}_t|S_t = k)\sum_l \alpha_l(t-1)r_{l,k} \quad (6)$$

The verification process shown in Algorithm 1 works as follows. The variable $flag$ records the test result from current

subsequence. If $flag == Un$ it means that there is no decision from current subsequence, and we increase the length of the sequence and continue the test. During the process of testing each subsequence, we calculate the probability of the sequence based on the genuine model ($j == 0$) and background model ($j == 1$), respectively. Also, we maintain the forward probability matrix in each test, and when increasing the length of the sequence by one we only have to update the forward probability matrix given the current length.

4 Experiments

In this section, we present the experimental results of our mobile user verification method. We test the user verification in both positive verification (p-verify) and negative verification (n-verify). In the former case, we use one user's own unseen data to test against his/her profile model and validation model. In the latter case, we use all the other users' data to attack the current user's model.

Data set. We use the mobility data recorded by the Device Analyzer app [Wagner *et al.*, 2014; 2013]. In this data set, all the location relevant and identifiable information is anonymized, such as cell tower IDs, WiFi AP mac addresses, etc., and is done separately by each user. Randomly choosing a testing period from 2013-01-01 to 2013-03-31, we select the candidate users in our experiments based on the following criteria. First, the selected users should have frequent cell tower information and/or WiFi information. Second, in order to have the balanced data set, each individual should have at least 10 one-week segments. According to these criteria, two data sets are obtained. In the first data set (D_1), we have 59 users from time zone "+0", and a total of 786 one-week data segments are accumulated. In the second data set (D_2), we have 50 users from time zone "+1", and a total of 636 one-week data segments are accumulated.

Parameter setting. In the training phase, we use each user's first 9 weeks' data to train the HMM, and the number of states is tested from 3 to 6. In the validation phase, we set the number of similar models and dissimilar models as the default value 5. This is because when applying the max function on the results from the background model lists, increasing the number of similar models or dissimilar models would not affect the results since it always uses the one with highest probability from the lists. The verification process is conducted on each user's remaining 2 to 4 weeks' data.

Evaluation metric. Since in each user's case the percentage of n-verify (98%) in terms of detection of the imposters is much larger than p-verify (2%), we evaluate the results in terms of three measures, namely, sensitivity, specificity, and undetermined rate in p-verify. According to the information given in Table 1, we calculate these measures given as follows

$$sensitivity = \frac{N_{tp}}{N_{tp} + N_{fn} + N^p_{un}}$$

$$specificity = \frac{N_{tn}}{N_{fp} + N_{tn} + N^n_{un}}$$

$$undetermined \quad rate, P_{un} = \frac{N^p_{un}}{N_{fp} + N_{tn} + N^p_{un}}$$

Table 1: Verification measures. Besides traditional terminologies we used, e.g., the number of verifications given true positive results (N_{tp}), the number of verifications given false positive results (N_{fp}), etc., we also have two additional quantities, namely, the number of undetermined verifications among the positive tests (N^p_{un}), and the number of undetermined verifications among the negative tests N^n_{un}.

		p-verify	n-verify
test results	accept H_0	N_{tp}	N_{fp}
	reject H_0	N_{fn}	N_{tn}
	undetermined	N^p_{un}	N^n_{un}

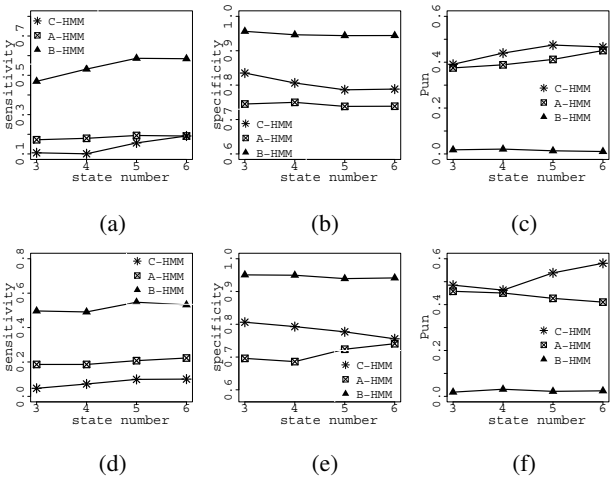


Figure 3: The verification results based on data set D_1 are shown in (a), (b), and (c), and each of them corresponds to the metrics of sensitivity, specificity, and undetermined rate in p-verify, respectively. The similar results obtained from D_2 are shown in (d), (e) and (f).

We find in our experiments that the undetermined rates in n-verify according to the different methods are typically very minor compared with those in the p-verify. Therefore, in the following sections, we do not evaluate the undetermined rate in the n-verify. This will be detailed in the follow-up section.

4.1 Comparison with other methods

We conduct the verification based on each user one-day's data according to Algorithm 1. To the best of our knowledge, there are few studies targeting on user verification using location data. Thus, the comparison is made among the proposed method using either one type of the observation, namely, C-HMM, A-HMM, or both types of the observations, namely, B-HMM.

The comparison results based on D_1 and D_2 are shown in Figure 3. A few observations can be made as follows.

First, in terms of sensitivity shown in Figure 3(a) and (d), the methods of using only one type of the observation, i.e.,C-HMM, A-HMM, perform unsatisfactorily, and in most cases the achieved results are less than 20%. Comparatively, the performance of using both types of the observation, i.e., B-HMM, is are much better, and in the best case the verification

accuracy in p-verify, i.e., the sensitivity, can be close to 60% given the number of hidden states is 6. This shows that, each type of the location data alone is unable to describe users' regular mobility behaviors. And these two types of data are complementary in the scenario of modeling users' mobility patterns since the verification results are greatly improved via combining these two types of data.

Second, similar to the results of sensitivity, Figure 3 (b) and (e) show that combining two types of data can greatly improve the specificity than those using one type of data. Also, in the best case, B-HMM can achieve around 95% verification accuracy in n-verify, i.e., specificity, which makes our method applicable in the real case of user verification. It is worth to note that, the great difference between the achieved specificity (around 95%) and sensitivity (around 60%) given different methods indicates that the main challenge in user verification using location data is how to recognize users' own mobility behaviors, i.e., p-verify in our case, rather than detecting the different mobility behaviors from others, i.e., n-verify in our case.

Third, due to the high specificity achieved in different methods, the corresponding undetermined rate is small in n-verify. Therefore, we only show the plot of the undetermined rate in p-verify, P_{un}, given in Figure 3 (c) and (f). The results show that given two types of data we can almost make a decision of 99% of the one-day sequences, where among the p-verify we can achieve 58% accuracy and among the n-verify we can achieve 94% accuracy by using B-HMM.

4.2 Other factors in user verification

In this section, we explore several other facts that can affect user verification results, e.g., the starting hour of the test sequences, and the minimum length of the sequences used in verification. Note that in the training phase we fix the number of hidden states to be 6.

Figure 4 (a) and (b) show the achieved accuracy in two types of verifications in different cases. Given different starting hour, the verification accuracy increases with the length of the sequence increasing in the p-verify, and it shows an opposite trend in the n-verify. The former one is consistent with the finding in the previous section that, short period data may not be sufficient to describe users' regular mobility behaviors, and increasing the length of the testing sequence will greatly improve the results. The latter one is due to the reason that long sequence may include certain overlapped mobility behaviors, e.g., staying in the office, and it leads to a lower verification accuracy when the length of the sequence increases. However, the effect of increasing the length of the mobility sequence in user verification is minor since the accuracy in n-verify decreases only 1% when increasing the length of the sequence from 2 to 12. This result further validates that, starting with a short period data, e.g., 2 hours' data, can achieve a reasonably high verification accuracy, around 94%, when the smartphone is being used by those other than the owner.

The temporal information also affects the verification results. When the minimum length of the sequence is short, e.g., less than 6 hours, the best starting hour in confirming the current user is owner is 12:00, and given longer sequence the best starting hour is 17:00 shown in Figure 4 (a). When detecting the abnormal usage of the smartphones, the sequences starting from 17:00 are always a good choice shown in Figure 4 (b). This is because the sequence starts from 17:00 may include one user's typical daily route, namely, a short period of staying at office, the transition period from office to home, and the period of staying at home.

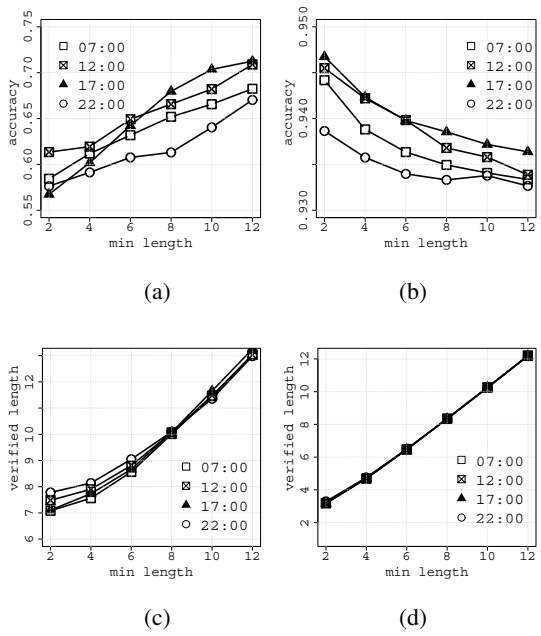


Figure 4: The verification accuracy of the p-verify (a) and n-verify (b) based on different minimum length and starting hour in D_1. The mean length of the correctly verified sequence in the p-verify (c) and n-verify (d) based on different minimum length and starting hour in D_1.

Figure 4 (c) and (d) further show the mean length of the sequence used in the correct p-verify and n-verify. In the p-verify, when we set the minimum length of the test sequence to be 2, it takes on average 7 to 8 hours' data to determine that a given user is the legitimate user. Comparatively, in the n-verify, the mean length of the sequence used always a little bit longer, e.g., one hour, than the minimum length.

5 Conclusion

In this paper, we present a novel user verification method by combining the complementary location ambience features on location anonymized data set. The proposed method shows a high detection accuracy of 94% for illegitimate users by using two hours' data and a verification accuracy of 72% with less than one day's data when the smartphone is used by its owner. The proposed method can be applied in complementing and enhancing current smartphone user authentication towards less imposition of user's load. Our future work along this avenue includes the investigation of finding better mobility features and model user's mobility with an aim to further reduce the false alarm rate and to improve the useability of the passive user verification method.

Acknowledgments

This work is supported by Ministry of National Development (MND) Singapore under the grant No. SUL2013-5. And it is also supported by the research grant for the Human-centered Cyber-physical Systems Programme at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology and Research (A*STAR).

References

[Agrawal and Srikant, 1995] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In *11th International Conference on Data Engineering*, pages 3–14, 1995.

[Bayir *et al.*, 2009] M.A. Bayir, M. Demirbas, and N. Eagle. Discovering spatiotemporal mobility profiles of cellphone users. In *10th IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks*, pages 1–9, 2009.

[Bimbot *et al.*, 2004] Frédéric Bimbot, Jean-François Bonastre, Corinne Fredouille, Guillaume Gravier, Ivan Magrin-Chagnolleau, Sylvain Meignier, Teva Merlin, Javier Ortega-García, Dijana Petrovska-Delacrétaz, and Douglas A. Reynolds. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, 2004:430–451, 2004.

[de Montjoye *et al.*, 2013] Yves-Alexandre de Montjoye, Cesar A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, 3, 2013.

[De Mulder *et al.*, 2008] Yoni De Mulder, George Danezis, Lejla Batina, and Bart Preneel. Identification via location-profiling in gsm networks. 7th ACM Workshop on Privacy in the Electronic Society, pages 23–32, 2008.

[Dempster *et al.*, 1977] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society, series B*, 39(1):1–38, 1977.

[Do, 2003] M.N. Do. Fast approximation of kullback-leibler distance for dependence trees and hidden markov models. *Signal Processing Letters, IEEE*, 10(4):115–118, 2003.

[Gambs *et al.*, 2013] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Nunez Del Prado Cortez. De-anonymization attack on geolocated datasets. In *12th IEEE Inter. Conf. on Trust, Security and Privacy in Compt. and Communications*, pages 789–797, 2013.

[Kullback and Leibler, 1951] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 03 1951.

[Lin *et al.*, 2015] Miao Lin, Hong Cao, Vincent Zheng, Kevin C. Chang, and Shonali Krishnaswamy. Mobile user verification/identification using statistical mobility profile. In *Second International Conference on Big Data and Smart Computing*, pages 15–18, 2015.

[Sahoo *et al.*, 2012] Soyuj Kumar Sahoo, Tarun Choubisa, and S. R. Mahadeva Prasanna. Multimodal biometric person authentication : A review. *IETE Technical Review*, 29:54–75, 2012.

[Song *et al.*, 2010a] Chaoming Song, Tal Koren, Pu Wang, and Albert-Lázló Barabái. Modelling the scaling properties of human mobility. *Nature Physics*, 6:818–823, 2010.

[Song *et al.*, 2010b] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-Lázló Barabái. Limits of predictability in human mobility. *Science*, 327:1018–1021, 2010.

[von Zezschwitz *et al.*, 2013] Emanuel von Zezschwitz, Paul Dunphy, and Alexander De Luca. Patterns in the wild: A field study of the usability of pattern and pin-based authentication on mobile devices. In *15th International Conference on Human-computer Interaction with Mobile Devices and Services*, pages 261–270, 2013.

[Wagner *et al.*, 2013] Daniel Wagner, Andrew Rice, and Alastair Beresford. Device analyzer: Understanding smartphone usage. In *10th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, 2013.

[Wagner *et al.*, 2014] Daniel T. Wagner, Andrew Rice, and Alastair R. Beresford. Device analyzer: Largescale mobile data collection. *SIGMETRICS Perform. Eval. Rev.*, 41(4):53–56, 2014.

[Wald, 1945] A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 06 1945.

[Zheng *et al.*, 2008] Vincent Wenchen Zheng, Evan Wei Xiang, Qiang Yang, and Dou Shen. Transferring localization models over time. In *23rd AAAI Conference on Artificial Intelligence*, pages 1421–1426, 2008.