Identification via Location-Profiling in GSM Networks

Yoni De Mulder
Katholieke Universiteit Leuven
Dept. Elect. Eng.-ESAT/SCD-COSIC
Kasteelpark Arenberg 10, 3001 Heverlee, Belgium
yoni.demulder@esat.kuleuven.be

George Danezis
Microsoft Research
Cambridge, UK
gdane@microsoft.com

Lejla Batina
Katholieke Universiteit Leuven
Dept. Elect. Eng.-ESAT/SCD-COSIC
Kasteelpark Arenberg 10, 3001 Heverlee, Belgium
lejla.batina@esat.kuleuven.be

Bart Preneel
Katholieke Universiteit Leuven
Dept. Elect. Eng.-ESAT/SCD-COSIC
Kasteelpark Arenberg 10, 3001 Heverlee, Belgium
bart.preneel@esat.kuleuven.be

ABSTRACT

As devices move within a cellular network, they register their new location with cell base stations to allow for the correct forwarding of data. We show it is possible to identify a mobile user from these records and a pre-existing location profile, based on previous movement. Two different identification processes are studied, and their performances are evaluated on real cell location traces. The best of those allows for the identification of around 80% of users. We also study the misidentified users and characterise them using hierarchical clustering techniques. Our findings highlight the difficulty of anonymizing location data, and firmly establish they are personally identifiable.

Categories and Subject Descriptors

K.4.1 [**Computers and Society**]: Public Policy Issues—*Privacy*; K.6.5 [**Management of Computing and Information Systems**]: Security and Protection

General Terms

Performance, security, experimentation

Keywords

Identification, location profile, cellular network, location privacy

1. INTRODUCTION

Mobile phone devices have become indispensable items of everyday life, since the GSM and cellular technologies were introduced and became popular in the 1990s. Reliance on them, to support business as well as leisure, is likely to increase with the current deployment of third generation services providing high-bandwidth data services, as well as location based services.

In order to function, and route calls, these technologies require the service provider to know the cell in which a mobile device is present. These cells are of varying size, from a few kilometres in low-density areas, to a few meters within cities. This gives service providers a record of the movement of each device, and probably its owner. This represent a serious privacy threat, and previous research has explored the public perception surrounding it [4].

Records of devices movements are routinely kept by service providers, and often used as part of investigations by law-enforcement. The data retention directive has been implemented in some EU countries to mandate the retention of location data, providing anyone with access to the data a map of all mobile devices movements, at a cell granularity [5].

In this work we assess the degree to which these records are personally identifiable information. In particular we assess the extent to which anonymised location records from cell based mobile phone networks can be linked back to previously extracted user profiles. We show techniques to build profiles based on some user's locations, and then techniques to match those profiles with anonymised location data.

Our approaches are evaluated using the real-world location traces of mobile users from the MIT Reality mining project [1]. This demonstrates that our techniques are robust to noise and artefacts present in real-world data, and would perform well in live conditions. Our key measures of success are the rates of correct identification of the anonymised traces. In cases identification fails, we still manage to cluster and order users according to how likely they are to be the user behind the anonymised trace. We show that in most cases the deanonymisation either succeeds or leads to only a handful of candidate users.

Our results have a profound impact on how we perceive cell location data. First, some location privacy approaches rely on changing mobile identifiers quite often [2]. We show

WPES'08, October 27, 2008, Alexandria, Virginia, USA.

that the periods have to be extremely short, since it is easy to identify a moving user within hours.

Second we demonstrate that it is difficult, if not impossible to sanitise and anonymise location data, merely by removing user identifiers and reducing the granularity of the location or time. Cell locations in fact blur the exact location of users, but a sequence of cells allows us to identify users with a very high probability. Even sanitised location data is still personally identifiable.

Finally naive approaches to protecting one's privacy, such as using two different pre-paid mobile phones, or changing mobile phone, are not likely to provide any protection against a determined adversary. Given just the cell locations of the devices it will be trivial to infer that they belong to the same person, unless the actual places visited, places of work and stay of the user change at the same time – this is a rather extreme precaution to take merely to achieve some privacy.

This paper is organised as follows: we first give a brief overview of mobile phone network architectures, and how locations of mobiles are advertised to the service providers in section 2. In section 3 we present how to extract location profiles for users from movement logs. Those profiles are then used to identify further traces of movements, as presented in section 4. In case identification fails we extract possible groups within which a user may be by clustering (section 5). We evaluate our approach on a live data set and present out results in section 6. Finally we offer some conclusions as well as avenues for future research in section 7.

2. CELLULAR NETWORKS

A cellular or GSM network consists of cells covering a certain area served by a fixed base station. The cell shape of real cellular networks varies depending on the base station antenna radiation pattern and each cell can have a number of neighbouring cells. Each base station is connected to a Mobile Switching Center (MSC), which is, in turn, connected to the Public Switched Telephone Network (PSTN). All communications are mediated through the base station, and mobile stations (i.e., mobile phones) talk with one another via a base station, not directly. In general, a cellular network can be referred to as a *zone*-based network. A zone can be a Location Area (LA), as in current GSM cellular networks, or a cell depending on the system.

The Location Management (LM) functionality of the network finds out the cell of a mobile station in order to route incoming calls in an efficient manner. Location management involves two operations: location updates and paging. A location update is sent by the mobile station to let the cellular network know its current position. The mobile registers its location with the base station, and the location databases are modified accordingly. Paging, on the other hand, is performed by the network to find the cell in which a mobile station is located. Only location updates are of importance to this work.

Location update schemes can be *static* or *dynamic*. In static location update schemes the mobile stations advertise their locations without taking into account any other user characteristics. Some simple schemes include the *Always-Update* and *Never-Update*, that always broadcast the mobile's location when it moves between cells or never do so respectively. Dynamic location update schemes are defined according to each individual user. Threshold-based schemes trigger a location update periodically or when the mobile crosses a certain number of cells, or moves a certain distance. It can also be triggered by the frequency of incoming calls. Profile-based location updates only require updates when the user is outside a certain *home* area, according to a profile extracted by the network operator [6]. These profiles are likely to be very similar to what we use for our identification procedures, and should match the user's movement behaviour.

3. MOBILITY MODEL AND LOCATION PROFILES

3.1 Previous work on mobility models

We consider a *cell-based* GSM cellular network with a *static always-update* location update scheme. In such a network each mobile user registers their location with the base station of the current cell at each cell boundary crossing. Consequently, the network obtains a sequence of cell-ID's. The static always-update cell-based location scheme captures mobile users' movement in their finest details and the network has full knowledge of the location of users. We use these sequences of cell-IDs per user to build profiles, as well as to identify anonymous mobile devices.

The definition of movement history mentioned in [3] is applied: The movement history of a mobile user is a string '$v_1 \cdot v_2 \cdot v_3 \cdot \ldots$' of symbols of the alphabet ϑ, where ϑ is the set of zones in the cellular network and hence v_i denotes the zone-ID reported by the i^{th} location update. In our approach, ϑ is the set of all cells in the network. Since the update scheme is static always-update, the successive cell-ID's v_i are distinct and in most cases neighbouring cells.

It is important to note that these traces could be collected in many ways. Obviously they are available to operators in real-time as location updates are received. These location updates can, and are routinely, logged to optimise the usage of the network as well as under regimes of traffic data retention for law enforcement purposes. The traces can also be acquired by tampering with the mobile device of a user, so that it records visited cell locations. An eavesdropper can attach to a user (or vehicle) under surveillance a mobile device that either transmits or records cell locations. Finally profiles could be extracted by sampling a user's location through physical or CCTV surveillance, but such processes are outside the scope of this work.

Mobility models make use of past movement traces to build probabilistic models and predict the mobile user's future locations. Their key assumption is that users have some set locations, and movement patterns that are likely to not significantly change over time. User mobility models traditionally play an important role in designing location update schemes, and optimizing network use. In our case, we use mobility models to identify users. Two mobility models have been discussed in previous work [8], [3]:

- **Random Walk:** This model is regularly used to model the movements of users in a cellular network. It assumes that the direction of movement is random, and hence users visit each neighbouring cell with equal probability. It only requires the current location to predict the next cell occupied by a user and as such can be seen as a memory-less movement model.

- **Markovian:** In contrast to the previous model, this approach uses a per-user profile. The Markovian mobility model defines distinct probabilities for movement from a given cell (or sequence of cells) to each of its neighbours (or the last cell in the sequence). These transition probabilities are dependent on individual user movement histories.

  Markovian mobility models can have a different order, indicating the number of cells to take into consideration when calculating the transition probabilities. In the simplest case, transition probabilities represent the probabilities of moving from a single cell to one of its neighbouring cells. These probabilities are referred to as *one-step transition probabilities* and are modelled as an *order-1 Markov chain*.

In [6], the continuous-time order-1 Markovian mobility model was used for a dynamic predictive location management scheme in order to reduce the combined location updating and paging cost. Our model will be based on a similar Markovian model, but our objective is different. The location profiles based on the model, are used to perform identification in a cellular network of a mobile user based on the location profiles of all users.

3.2 Location Profiles

In the *continuous-time order-1 Markovian mobility model*, cells are treated as states of a Markov chain, and each cell change corresponds to a state transition. A cell change of a mobile user can occur at any time. At some time, a mobile user moves into cell i. After it spends a dwell[1] time in cell i, it will move to one of its neighbouring cells, e.g. cell j, with a *one-step probability* $p_{i,j}$. Those probabilities form the Transition Probability Matrix described below.

For each mobile station in the GSM cellular network, a separate location profile is maintained. This profile is based only on the cell movement history of the mobile user. The Markovian profile of each user consists of the transition matrix and its stationary distribution.

- The *Transition Probability Matrix* (TPM) takes a mobile user's movement behaviour and geographical factors (neighbouring cells) into consideration. For a mobile user who resides within m cells, the TPM $\mathbf{P} = ((p_{i,j}))$ is an $m \times m$ matrix.
  The element $p_{i,j} = \Pr(\text{Cell}_j|\text{Cell}_i)$,$(i, j = 1, \ldots, m)$ of the TPM is the probability with which a mobile user will move to cell j upon leaving current cell i, hence $0 \leq p_{i,j} \leq 1$. The sum of all probabilities with which a mobile user moves to all possible (neighbouring) cells j upon leaving cell i is 1; $\sum_{j=1}^{m} p_{i,j} = 1$ for $i = 1, \ldots, m$. $p_{i,j}$ depends on the current cell i, the next cell j and the mobile user's movement.

  TPMs for individual mobile users can be estimated from a set of location traces. Assuming the location update scheme is *static always-update*, the mobile device always updates its location on cell boundary-crossing and never performs a location update when staying in the current cell. Thus $p_{i,i} = 0$ for $i = 1, \ldots, m$. The one-step transition probabilities $p_{i,j}$ are computed by the relative counts "$\text{Count}(\text{Cell}_i \rightarrow \text{Cell}_j)$". Then $p_{i,j}$ corresponds with the number of cell transitions $\text{Count}(\text{Cell}_i \rightarrow \text{Cell}_j)$ divided by the total number of cell transitions out of Cell_i, as shown in:

$$p_{i,j} = \frac{\text{Count}(\text{Cell}_i \rightarrow \text{Cell}_j)}{\sum_{j=1}^{m} \text{Count}(\text{Cell}_i \rightarrow \text{Cell}_j)} \quad (1)$$

- The *stationary Markov distribution* $\mathbf{\Pi}$ of the markov chain represents the user residence probabilities in each cell of the cellular network. For a mobile user who resides within m cells, $\mathbf{\Pi}$ will be a $m \times 1$ vector: $\mathbf{\Pi} = [\pi_1 \cdots \pi_m]^T$. The element $\pi_i = \Pr(\text{Cell}_i), (i = 1, \ldots, m)$ of $\mathbf{\Pi}$ is the residence probability of a mobile user in cell i. The sum of all user's residence probabilities is 1; $\sum_{i=1}^{m} \pi_i = 1$.

  This distribution, also called the steady-state probability vector $\mathbf{\Pi}$, can be easily computed using the Transition Probability Matrix $\mathbf{P}$ by solving $\mathbf{\Pi} = \mathbf{\Pi} \times \mathbf{P}$, and hence all stationary Markov distributions $\mathbf{\Pi}$ are user-dependent as well. A more practical way to compute $\mathbf{\Pi}$ is $\lim_{k \rightarrow \infty} \mathbf{P}^k = \mathbf{\Pi}$, where in practice $k = 20$ will suffice.

4. IDENTIFICATION PROCESSES

The main objective of this paper is to assess how identifiable a mobile user is in a cellular network, e.g. GSM network, based on his movements within the network, and some known location profiles for the user population.

The full identification process is based on processing data in two distinct periods.

- A single, possibly anonymous, user is observed and his location information is recorded from location messages for a certain time span, referred to as the *identification period*. (We consider an identification period of up to one month.)

- The movements during the identification period are compared with the identification database containing profiles based on a month preceding the *identification period*. This comparison yields an *identification indicator* for each possible user profile. Based on these *identification indicators*, the user profile most likely to correspond to the target user is selected.

Two different identification processes are presented and their performances are evaluated in section 6. They both make use of distinct first order markovian location profiles for users, describing their past movements. Their key difference is the way they match the location information from the identification period to those profiles.

4.1 Identification process based on Markovian model

A new location profile is generated from the location update messages sent from the device of mobile user U_x during the identification period. The new profile is compared for closeness with the profiles stored in the identification database. The new profile for the target consists of a Transition Probability Matrix $\mathbf{P_x}$ and a stationary Markov distribution $\mathbf{\Pi_x}$.

Once the location profile of mobile user U_x is generated, it is compared with all location profiles present in the identification database. For each mobile user $U_k, (k = 1, \ldots)$ of the

[1]Time interval between successive cell changes i.e. residence duration.

user population in the database, an *identification-indicator* $iden_k$ is calculated:

$$iden_k = \sum_{i,j}^{n} Pr_x(Cell_j|Cell_i) \cdot Pr_x(Cell_i) \cdot Pr_k(Cell_j|Cell_i) \cdot Pr_k(Cell_i) \quad (2)$$

$$= \sum_{i,j}^{n} p_{x;i,j} \cdot \pi_{x;i} \cdot p_{k;i,j} \cdot \pi_{k;i} \quad (3)$$

where n denotes the number of all possible cells in the cellular network used in this research but in practice only the cells in which mobile users U_x and U_k reside matter. The identification index $iden_k$ compares both transition matrices and stationary Markov distributions of mobile users U_x and U_k.

The mobile user U_{iden} identified by the process corresponds to the mobile user U_k with the largest value of $iden_k$:

$$U_{iden} = \arg \max_{k \in \{1,\dots\}} iden_k \quad (4)$$

The higher the value of $iden_k$, the better profiles resemble, meaning that the mobile users U_x and U_k have the same cell-behaviour within the cellular network.

The performance of this identification process is evaluated in section 6.2.

4.2 Identification process based on sequence of cell-ID's

The location information gathered during the identification period is a chronological sequence of cell base stations in the form of tower IDs. This process evaluates the likelihood this sequence was generated by the different location profiles stored in the identification database.

In this identification process, a sequence number θ is assigned to each transition of the sequence of cell tower IDs observed during the identification period. If the location sequence consists of l cell tower IDs, it contains $l-1$ cell transitions and hence $\theta = 1, \dots, l-1$. To compute $iden_k$, the one-step transition probability is looked up for each transition θ in the transition probability matrix $\mathbf{P_k}$ of location profile U_k[2]. We denote this probability by p_k^θ. The indicator $iden_k$ corresponds with the product of all these transition probabilities, given by following formula:

$$iden_k = \prod_{\theta=1}^{l-1} p_k^\theta \quad (5)$$

where l denotes the number of cell tower IDs present in the location sequence. In case one of the transitions have a zero probability we assign to it a very small probability (e.g. 10^{-8}) to ensure that we never get $iden_k = 0$.

Because each transition probability $p_k^\theta \in [0,1]$, the product of them ($iden_k$) quickly becomes very small. Consequently, we use the logarithm of the product of transition probabilities to calculate $iden_k$, which is equivalent to the sum of the logarithm of the probabilities:

$$iden_k = \log_{10}(\prod_{\theta=1}^{l-1} p_k^\theta) = \sum_{\theta=1}^{l-1} \log_{10} p_k^\theta \quad (6)$$

[2]Only the Transition Probability Matrix is used, and not the stationary Markovian distribution.

The indicator $iden_k$ will always be negative because the logarithm is negative within $[0,1]$. As mentioned earlier, it is possible to stop the calculation of $iden_k$ earlier and look at the intermediate values of all $iden_k$. The mobile user with the highest value of $iden_k$ (see formula (4)) is selected as they have the highest cell-resemblance to mobile user U_x.

In contrast to the previous process, it is easy to vary the length of the *identification period* so the performance of this identification process will be examined for four identification periods: one hour, one day, one week and month. The results discussed in section 6.3 give some insight into the correlation between the length of the *identification period* and the identification performance.

5. CLUSTERING

5.1 Purpose of clusters

We use a clustering algorithm to group together mobile users based on their cellular behaviour, or more precisely their cell residence probabilities. Clusters are generated among the mobile user profiles in the identification database based on the stationary Markovian distributions.

Clustering is used to evaluate the identification process and to understand the nature of false positives, when an incorrect mobile user is identified. First of all, we can examine whether the original mobile user[3] is part of the same cluster as the incorrectly identified mobile user. In those cases, the identification process is expected to perform poorly, since both mobile users show a similar cellular behaviour.

The percentage of these clustered incorrectly identified users can be used as a performance measure of the identification processes: the higher this percentage, the more expected the process to fail. Secondly, the probability density estimate function of the cosine similarity (explained 5.2) between the *original user* and the corresponding user in the identification database (the one that should have been identified) and of the cosine similarity between the *original user* and the *wrongly identified user* out of the database is examined. The distance between these two pdf's[4] can also be taken as a performance measure of the identification process: if this distance is small, the cosine similarity of the incorrect identified users is close to those of the original 'to be identified' users, and thus the identification process is expected to fail and to perform poorly.

Our main goal is to keep the percentage of incorrectly identified mobile users as small as possible. But when false positives occur, it is acceptable to have errors when the distances are small and hence a large clustered group. All these conclusions are discussed based on the results of the identification processes in section 6.

In the following section, we present an outline of the algorithms involved in our cluster algorithm.

5.2 Agglomerative Hierarchical Clustering

Vector space model. Clusters are formed among mobile users based on their cell residence probabilities, described by the stationary Markovian distributions $\mathbf{\Pi}$ already computed in the location profiles of the identification database. In our

[3]The mobile user out of the database corresponding to the *original user*

[4]This equals the difference in x-values corresponding with the pdfs' maxima.

case, the vector space model corresponds with the set of stationary Markovian distributions $\mathbf{\Pi}$ of all mobile users in our user population. This model will form the basis of the cluster formation algorithm.

Distance matrix D. Any clustering algorithm requires a *distance metric* to assess the resemblance or distance between two different objects. In our case we need a measure of the resemblance in cellular behaviour between two mobile users. The more similar cellular movement and residence the users show, the smaller the distance between them. For this purpose a *symmetric distance metric* is applied, i.e. the *cosine distance*. The cosine distance equals one minus the cosine similarity between two vectors. This is the cosine of the angle between the stationary Markovian distributions $\mathbf{\Pi}$ of both users. If the users have a similar cellular behaviour and residence, the angle between the vectors is small and the cosine approaches one, and hence the cosine distance approaches zero.

In order to form clusters among the mobile users in the user population, the cosine distance between every possible pair of users is calculated. All these cosine distances are stored and organised in a *distance matrix* $\mathbf{D}$. Since the cosine distance metric is symmetric, so is the distance matrix.

Hierarchy of clusters. Based on the computed symmetric distance matrix $\mathbf{D}$ a hierarchy of clusters among the mobile users is created. The *agglomerative hierarchical clustering algorithm* is applied to assign each mobile user to a final cluster.

Agglomerative means that initially all n mobile users form a cluster on their own. Then clusters are aggregated until the final state is reached, when all mobile users belong to a single cluster. At each step a pair of clusters are merged based on their mutual distance. The *hierarchy of clusters* consists of the sets of clusters at each step of the algorithm.

Several methods can be used to select the clusters to merge. They differ in the manner they compute mutual distance between clusters. We use *average linkage* as a cluster distance for our algorithm: the mutual distance between two clusters depends on the cosine distances between all pairs of mobile users present in both clusters. At each step, the clusters with the smallest mutual distance are merged. The *hierarchy of clusters* can easily be presented by a *dendrogram* which involves a tree structure.

Optimal number of clusters. Of course, one cluster including all mobile users is not the optimal final composition of clusters amongst users. For this reason we stop the merging process once cluster distance reaches an *optimal merge distance*. The *L-method* explained in [7] is used to calculate this distance. In a nutshell the *L-method* monitors the rate of increase of the merge distance between clusters at each step of clustering. It ends the process when quality of the clusters decreases meaning that clusters composed of different objects start being merged together.

6. PERFORMANCE EVALUATION

6.1 Dataset & Evaluation method

To evaluate the performance of the two identification processes, we used the Reality Mining dataset made available by the MIT Media Lab [1]. The dataset consists of the behaviour of one hundred human subjects at MIT during the 2004-2005 academic year over the course of nine months. It represents the largest mobile phone experiment attempted in academia. The data was collected using one hundred instrumented Nokia 6600 smart phones. The information collected, included call logs, Bluetooth devices in proximity, cell tower IDs, application usage and phone status. The generated data represents approximately 500000 hours of data on users' location, communication and device usage behaviour. The dataset has been anonymized and made available to the academic community. We acknowledge the fact that the dataset is limited to one hundred users, but the aim of our research is to show that identification is possible and should be considered as a privacy threat.

For our experiments we only need the data concerning users' location. This data consists of the entry and exit time in each cell for each mobile user in chronological order, including the cell tower ID. It can be interpreted as the information gathered from the location update messages sent on basis of a *static always-update scheme* in our presumed *cell-based* network. This is slightly different from the current GSM network implementations, which is a location area (LA)-based network with the static location area update scheme, but the principle of our work remains the same.

The identification performance of both identification processes discussed in section 4 is evaluated. During two different time periods known as the *identification periods*, namely 'December 2004' and 'January 2005', anonymised location data is captured from the mobile user who we want to identify (further referred to as *original mobile user*). These anonymised location traces are matched to the *identification database* consisting of the location profiles of all mobile users covering a time span of the month preceding the *identification process*, namely 'November 2004' and 'December 2004' respectively.

The evaluation of the correctness of the identification processes involves the calculation of the percentage of the original mobile users who are identified correctly. Profiles for some months are missing, due to missing data, and the evaluation excludes those profiles. We also look at the probability density estimate functions of the cosine similarity between the *original user* and the *correctly identified user* out of the database against all other non-identified users of the database. This cosine similarity is computed using the stationary Markovian distribution of the *original user* and those of the mobile users in the identification database.

In some cases, the identification process fails, and an incorrect mobile user is identified. It is expected that the identification process is misled when the original and the incorrect identified user have very similar movement profiles. Such users can be clustered together. As mentioned earlier in section 5.1, this can be used as a performance measure for an identification process. It is possible to look at the exact cosine distance between both the original user and the incorrectly identified user. The smaller this value, the more similar the users are.

It is also interesting to look at the probability density estimate functions of the cosine similarity between the *original user* and the corresponding user in the identification database (the one that should have been identified) and of the cosine similarity between the *original user* and the *identified user* out of the database. The distance between these two pdf's[5] can be taken as a performance measure as well. If this distance takes on small values, the cosine similarity of

[5]This equals the difference in x-values corresponding with the pdfs' maxima.

Table 1: Performance

Dec 2004	Jan 2005
38.89%	37.5%
$\Pr(\mathrm{ID}_{\mathrm{Jan2005}}\|\mathrm{ID}_{\mathrm{Dec2004}})$	48.276%

Table 2: Clustered incorrect identified users

Dec 2004	Jan 2005
88.636%	74.286%

the incorrect identified users is close to those of the original 'to be identified' users, and thus the identification process performs poorly but it is expected that the process fails. We see that a large part of the incorrectly identified mobile users is situated in the same cluster as the actual users.

6.2 Identification based on Markovian model

First we evaluate the performance of the Markovian identification process as described in detail in section 4.1.

6.2.1 Evaluation of correct identification

Table 1 gives an overview of the percentages of the correctly identified mobile users for both periods. In both cases, more than a third of the original users are correctly identified: $\Pr(\mathrm{ID}_{\mathrm{Dec2004}}) = 0.39$ and $\Pr(\mathrm{ID}_{\mathrm{Jan2005}}) = 0.38$.

A comparison can be made between the correctly identified users in both periods 'December 2004' and 'January 2005'. It is not always the case that a mobile user who is correctly identified during one period is also identified during another period. In general, 48.276% (also shown in table 1) of the correct identified mobile users are present in both periods: $\Pr(\mathrm{ID}_{\mathrm{Jan2005}}|\mathrm{ID}_{\mathrm{Dec2004}}) = 0.48$.

Figure 1 shows the probability density estimate functions of the cosine similarity between the original user and the correct identified user out of the database against all other users of the database for both time periods. The pdf[6] of the correct identified user is centered around high values of the cosine similarity, in contrast with the pdf of all other users which is mainly centered around very low values. This behavior is expected because the correct identified user normally is the one with the highest cosine similarity among all other users in the same user population. It can be clearly observed that the two pdf's have different shapes. The density function of the identified users has higher variance since fewer observations are available than for the function of the other users.

6.2.2 Evaluation of incorrect identification

Table 2 gives an overview of the percentage of incorrectly identified users who are situated in the same cluster as the original user. For both periods, three-quarters or more incorrectly classified users exhibit this characteristic. We call such misidentifications 'clustered', in contrast with wrongly identified users not belonging to the same cluster, that we call 'Non-clustered'.

We examine the cosine distance between the profile of wrongly identified users of those two groups, and the correct user. The cosine distance acts as a parameter for how close the mobile users resemble. Each pair of the clustered group is closely related and shows similar cell movements corresponding to the small cosine distance. In these cases it comes as no surprise that the identification process fails. The location profile of the identified user belonging to the non-clustered group, which covers fortunately less than a quarter of the wrongly identified users, differs significantly from the original user (corresponding to a larger cosine distance), thus the failure of the identification is unexpected. For this groups we expect a better identification process to lead to better identification results, as it is the case for the second identification process.

Table 3: Comparison probability density estimate functions

	Dec 2004	Jan 2005
$\mathrm{CS}_{\mathrm{intersect}}$	0.797	0.7736
$\Pr(\mathrm{CS}_{\mathrm{intersect}} < \mathrm{CS}_{\mathrm{Or}} < 1)$	0.5754	0.3942
$\Pr(\mathrm{CS}_{\mathrm{intersect}} < \mathrm{CS}_{\mathrm{Id}} < 1)$	0.3573	0.2165
pdf's distance	0.10832	0.10902

Figure 2 shows the probability density estimate function of the cosine similarity of the incorrect identified mobile user and the user who should have been identified as the original user, for both time periods. The former pdf is centered around smaller values of the cosine similarity than the latter one. The small distance between the two pdf's, shown in table 3, is explained by looking at table 1, table 2 and considering the cosine distances discussed above. In both time periods, approximately two thirds of the mobile users are incorrectly identified, of which three quarters or more are situated in the same clusters as their original user. Moreover for this clustered group the cosine distances with the original user take on small values[7]. It is clear this identification process has difficulties distinguishing mobile users with similar cellular behaviour, with the clustered group representing more than three quarters of the mis-identified users.

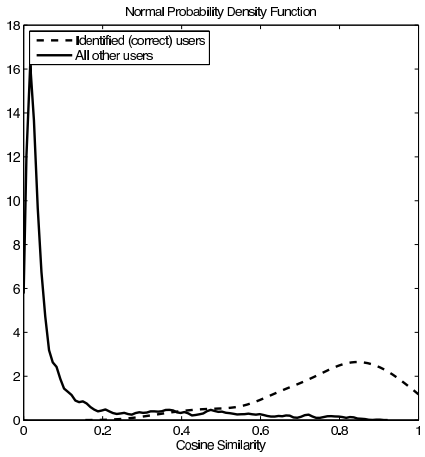
This observation can be strengthened when looking at the intersection of both pdf's, denoted by $\mathrm{CS}_{\mathrm{intersect}}$. The surface of the pdf between the values $\mathrm{CS}_{\mathrm{intersect}}$ and 1 for CS corresponds with the probability that CS lays in that interval: $\int_{\mathrm{CS}_{\mathrm{intersect}}}^{1} pdf(\mathrm{CS})d(\mathrm{CS}) = \Pr(\mathrm{CS}_{\mathrm{intersect}} < \mathrm{CS} < 1)$. In figure 2, surface '1' corresponds with $\Pr(\mathrm{CS}_{\mathrm{intersect}} < \mathrm{CS}_{\mathrm{Or}} < 1)$ and surface '2' with $\Pr(\mathrm{CS}_{\mathrm{intersect}} < \mathrm{CS}_{\mathrm{Id}} < 1)$. Table 3 gives an overview of these probabilities for both pdf's. The smaller the distance between the pdf's, the closer these probabilities are, the more overlapping the surfaces are and hence the less distinguishable the process is and the larger the clustered group within the false positives.

6.3 Identification based on sequence of cell-ID's

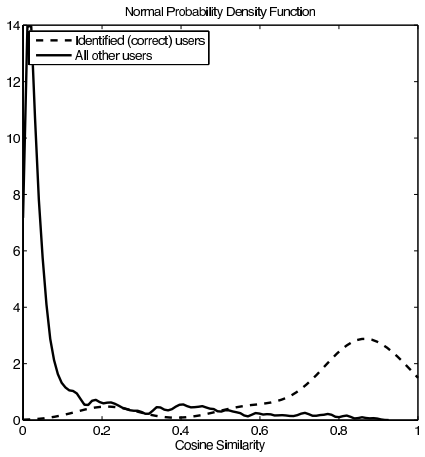
We evaluate our second identification process laid out in detail in section 4.2. For this process it is easy to vary the length of the *identification period*. Therefore we examine its performance for four different identification periods: one hour, one day, one week and a month. Increasing the identification periods results in better identification performances. The 'one month' period is discussed in detail for comparison with the previous identification process.

[6] probability density function

[7] which corresponds to large cosine similarity: $CS = 1 - CD$

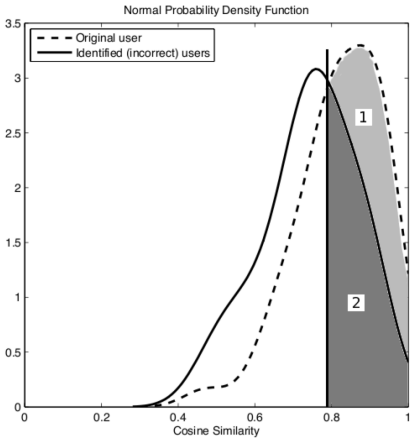


(a) Dec 2004

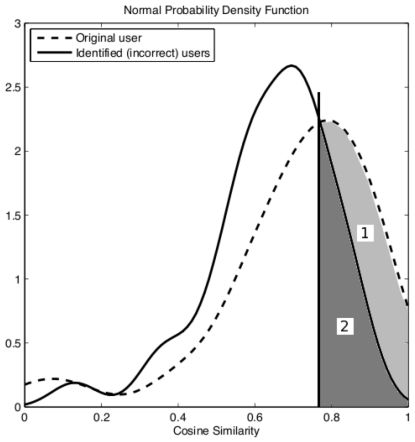


(b) Jan 2005

Figure 1: Probability density estimate function of cosine similarity of correct identified users versus all other users



(a) Dec 2004



(b) Jan 2005

Figure 2: Probability density estimate function of cosine similarity of incorrect identified users versus original user

Table 4: Performance

Dec 2004	**Jan 2005**
77.273%	87.755%
$\Pr(\mathrm{ID}_{\mathrm{Jan2005}}\|\mathrm{ID}_{\mathrm{Dec2004}})$	64.286%

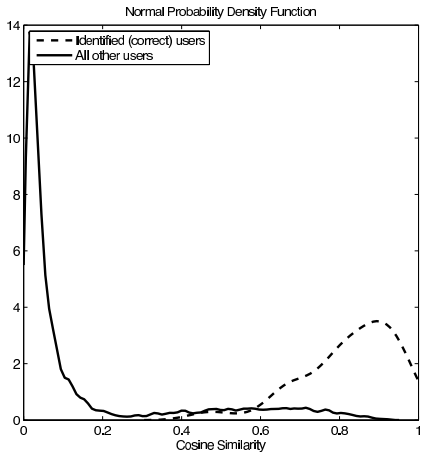
Table 5: Performance: all four identification periods

	Dec 2004	**Jan 2005**
one hour	40.91%	48.98%
one day	65.15%	65.31%
one week	74.24%	75.51%
one month	77.27%	87.76%

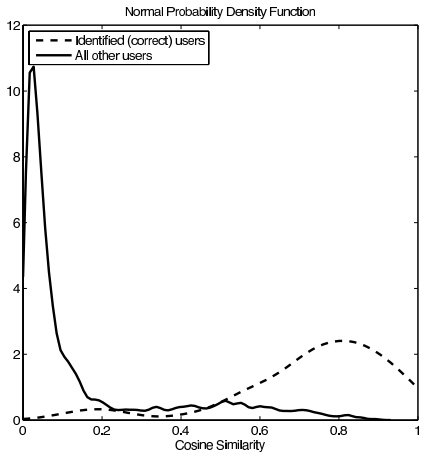
6.3.1 Evaluation of correct identification

As can be seen in table 4, more than three quarters of the original mobile users are properly identified: $\Pr(\mathrm{ID}_{\mathrm{Dec2004}}) = 0.77$ and $\Pr(\mathrm{ID}_{\mathrm{Jan2005}}) = 0.88$. Moreover, if one compares the correct identified users in both periods, there is an overlap of 64.286% (also shown in table 4):
$\Pr(\mathrm{ID}_{\mathrm{Jan2005}}|\mathrm{ID}_{\mathrm{Dec2004}}) = 0.64$. This greatly improves on the previous identification process.

Table 5 shows the identification performances for the four different periods in which anonymised location data of the original person is gathered. The longer the *identification period*, the more original users will be properly identified. If a performance of three quarters is sufficient, then an identification period of 'one week' is a good choice. There are diminishing returns in observing users for longer periods, since the probability of correct identification grows slowly.



(a) Dec 2004



(b) Jan 2005

Figure 3: Probability density estimate function of cosine similarity of correct identified users versus all other users

Table 6: Clustered incorrect identified users

Dec 2004	**Jan 2005**
60%	40%

Figure 3 shows the probability density estimate functions of the cosine similarity between the original user and the correctly identified user out of the database against all other users of the database for both time periods. The same remarks and conclusion can be made as mentioned in 6.2.1. In this case, the correctly identified user is the one with the highest cosine similarity among all other users in the same user population, and each correct identified user corresponds to all other non-identified users.

6.3.2 Evaluation of incorrect identification

Table 6 shows the percentage of the incorrect identified users who are situated in the same cluster as the original user. The percentages for both periods are smaller than those of the previous identification process. Fewer incorrect identified users show a similar cellular behavior with the original 'to be identified' user.

As before we can divide mis-identified users into clustered and non-clustered groups, and obtain the cosine distances between the real user and the mis-identified user. After comparison with the one of the previous identification process, the cosine distances have slightly larger values. This can be explained by the fact that the clustered group has become smaller and that the identification process performs better (it is capable of distinguishing better between mobile users with similar cellular behaviour.) As a result the similarity between incorrectly identified and original users is weaker.

Figure 4 shows the probability density estimate function of the cosine similarity of the incorrect identified mobile user and the one of the user who should have been identified corresponding to the original user for both time periods. Again the former pdf is centered around smaller values of the cosine similarity than the latter one, but the distance between

Table 7: Comparison probability density estimate functions

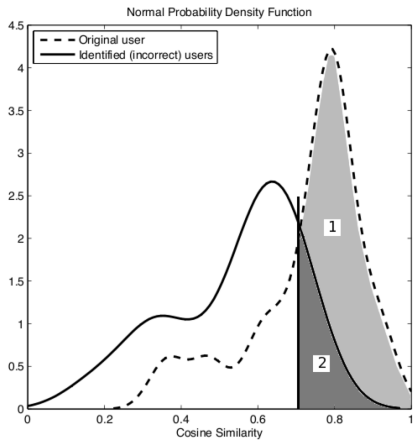
	Dec 2004	**Jan 2005**
$CS_{intersect}$	0.6996	0.7407
$Pr(CS_{intersect} < CS_{Or} < 1)$	0.687	0.379
$Pr(CS_{intersect} < CS_{Id} < 1)$	0.1926	0.1721
pdf's distance	0.15467	0.34562

both pdf's (shown in table 7) is slightly larger in comparison with figure 2 and table 3. This can be explained by looking at tables 4 and 6 and considering the cosine distance discussed above. In both periods, approximately one fifth of the mobile users are incorrectly identified. Out of those about half is not situated in the same cluster as the original user (non-clustered group), and moreover the corresponding cosine distances take on larger values. As mentioned earlier, this distance can be taken as a performance measure. Although this distance has increased relative to the previous process, the current identification process performs better since it distinguishes better between similar users' cellular behaviours who belonged to the clustered group of the previous identification process.

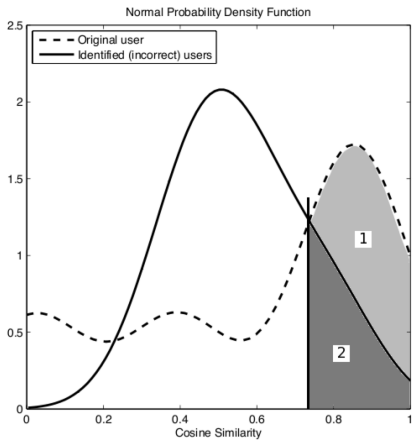
Our conclusions can be strengthened when looking at the intersection of both pdf's, denoted by $CS_{intersect}$. In figure 4, surface '1' corresponds with $Pr(CS_{intersect} < CS_{Or} < 1)$ and surface '2' with $Pr(CS_{intersect} < CS_{Id} < 1)$. Table 7 gives an overview of these probabilities for both pdf's. The larger the distance between the pdf's, the furhter these probabilities are, the less overlapping the surfaces are and hence the larger the non-clustered group for which the failure of identification is unexpected. This can be confirmed by comparing tables 3 and 7 and figures 2 and 4.

6.4 Performance comparison

Table 8 gives a summarized overview of the performance of both identification processes. Table 9 shows the percentage of the correct identified mobile users by the process based



(a) Dec 2004



(b) Jan 2005

Figure 4: Probability density estimate function of cosine similarity of incorrect identified users versus original user

Table 8: Performance comparison

Identification process based on ...	**Markovian model**		**Sequence of cell-ID's**	
Period of time	**Dec 2004**	**Jan 2005**	**Dec 2004**	**Jan 2005**
Performance (%)	38.89	37.5	77.273	87.755
Clustered incorrect identified users (%)	88.636	74.286	60.	40.
Unexpected group (%)	6.9445	16.0713	7.347	9.0908

Table 9: Comparison correct identified users

Dec 2004	**Jan 2005**
76.923%	88.235%

Table 10: Correct identified clustered users

Dec 2004	**Jan 2005**
71.795%	76.923%

on Markovian model who also are properly identified by the process based on sequence of cell-ID's. Table 10 shows the percentage of clustered mobile users that are incorrectly identified by the process based on the Markovian model but are properly identified by the process based on sequence of cell-ID's. The following observations can be made:

1. *The identification process based on the Markovian model performs poorly.* Slightly more than a third of the original mobile users can be correctly identified. Of the remainder, which involves the incorrect identified users, a large part (around 80%) is clustered with its original user due to similar cellular behavior and corresponding location profile.

2. *The identification process based on the sequence of cell-ID's performs well.* Around 80% of the original mobile users can be properly identified, and a smaller part (around 50%) of the incorrect identified users is clustered with its original user.

3. In both periods of time, around 80% of the users identified by the process based on the Markovian model will again be properly identified by the process based on the sequence of cell-ID's (table 9).

4. In both periods of time, around three quarters of the users that are incorrectly identified (and clustered) by the process based on the Markovian model, will be properly identified by the process based on the sequence of cell-ID's (table 10).

Combining these observations together, one can draw the following conclusion: The identification process based on the sequence of cell-ID's is more precise than the one based on Markovian model. The former process is able to distinguish better the incorrectly identified user from the original user when they are situated in the same cluster, in other words it distinguishes better mobile users with similar cellular behaviour. To be more precise, the former process identifies around 80% of the correct identified users by the latter process, and in addition properly identifies around three quarters of the clustered incorrect identified users by the latter process. This causes the percentage of clustered incorrect identified users to be smaller, which in turn causes the distance between the pdf's to increase.

We expect the better process to fail mainly on users that have highly clustered patterns of movement. This is not the case for the process based on sequence of cell ID's, which performs better, but has a smaller clustered group of mis-

identified users. One of the reasons for this discrepancy is the low number of false positives, that makes it difficult to evaluate whether users in the same cluster are more or less likely to be mis-identified than users in different clusters. At first sight it seems that the percentage of user mis-identified from the same cluster is lower than for the process based on Markovian model (with a large number of false positives) – a counter-intuitive result. This point requires further investigation.

Another reason for the lower fraction of clustered mis-identified users could be that the clustering is misleading: it takes into account similarity using the stationary distribution, which is a very naive metric compared with the probability of transition metric used by the best identification process. Maybe it is the similarity between users that is 'wrong' and not the identification process, i.e. for a more refined clustering distance metric the mis-identified users may actually appear far (not close). More work would be required to look at this.

When looking at table 1 and 5, one can clearly notice that the performance of the identification process based on the sequence of cell-ID's with the period 'one hour' already is better than the performance of the process based on Markovian model with the period 'one month'.

As mentioned earlier, the distance between the probability density functions of the incorrect and the corresponding original user can be taken as a performance measure. Although the main goal is to keep the number of false positives, i.e. incorrect identified mobile users, as small as possible, they do occur. In that case, an identification process is expected to perform poorly or to fail when a large part of the false positives belong to the clustered group.

When looking at the non-clustered group, where the process is not expected to fail, we want to keep this group as small as possible. The percentage of all mobile users who belong to this group, referred to as the *unexpected group*, can be computed as the percentage of non-clustered incorrect identified mobile users of the percentage of all incorrect identified mobile users: $\%_{\text{unexpected}} = \%_{\text{unclustered}} * \%_{\text{falsePositives}}$. Actually, the main goal is to minimize the number of false positives while minimizing in parallel the unexpected group as well. As can be clearly seen in table 8, the identification process based on sequence of cell-ID's is best at this task.

7. CONCLUSION

We have demonstrated that it is possible to build profiles of users' movements based on GSM location data, that make it possible to identify those users in a subsequent period with great accuracy (about 80% of the time.) The location profile models used are simple first-order markov chains. An obvious future avenue of research is to refine those models to increase the identification rate, by incorporating time or longer cell histories.

This work conclusively demonstrates that removing identifiers from location information, or merely blurring the spacial resolution, does not eliminate the danger or de-anonymization. It is likely that longer traces can still be mapped to profiles and re-identified. This has far reaching consequences for location privacy systems, as well as the protection that has to be afforded to stored location data as personally identifiable information.

These results also shed light on the way users price their location privacy, as studied in [4]. It was found that the payment sought to disclose twelve months of location information was much smaller (about twice) than the payment for a single month of data. Users may have an intuitive understanding that one month of location data already leaks the most important locations in their lives: their home, place of work or study, and their usual social spaces. It is rational to expect less payment for subsequent months, since the profiler learns little new information – in fact the information gathered in one month is so rich that it can be used to identify users for a long time after, as our results demonstrate.

8. ACKNOWLEDGEMENTS

The author Yoni De Mulder wishes to acknowledge the support and guidance of George Danezis and Lejla Batina as supervisors and Bart Preneel as promotor of his master thesis at COSIC at Katholieke Universiteit Leuven. He would like to express his thanks to George Danezis as well for helping with writing a paper out of his master thesis.

9. REFERENCES

[1] MIT Media Lab: Reality Mining. `http://reality.media.mit.edu/`, 2007.

[2] A. Beresford. *Location Privacy in Ubiquitous Computing*. PhD thesis, University of Cambridge, 2004.

[3] A. Bhattacharya and S. Das. LeZi-Update: An Information-Theoretic Framework for Personal Mobility Tracking in PCS Networks. *Wireless Networks*, 8(2):121–135, 2002.

[4] D. Cvrcek, M. Kumpost, V. Matyas, and G. Danezis. A study on the value of location privacy. In A. Juels and M. Winslett, editors, *WPES*, pages 109–118. ACM, 2006.

[5] C. Goemans and J. Dumortier. Enforcement issues – mandatory retention of traffic data in the eu: possible impact on privacy and on-line anonymity. *Digital Anonymity and the Law, series IT & Law*, pages 161–183, 2003.

[6] D. Gu and S. Rappaport. A dynamic location tracking strategy for mobile communicationsystems. *Vehicular Technology Conference, 1998. VTC 98. 48th IEEE*, 1, 1998.

[7] S. Salvador and P. Chan. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on*, pages 576–584, 2004.

[8] B. Sidhu and H. Singh. Location Management in Cellular Networks. In *Proceedings of World Academy of Science, Engineering and Technology*, pages 314–319, 2007.