How to Increase Security in Mobile Networks by Anomaly Detection

Roland Büschkes, Dogan Kesdogan, Peter Reichl
Aachen University of Technology - Department of Computer Science
Informatik 4 (Communication Systems)
D-52056 Aachen, Germany
{roland, dogan, peter}@i4.informatik.rwth-aachen.de

Abstract

The increasing complexity of cellular radio networks yields new demands concerning network security. Especially the task of detecting, repulsing and preventing abuse both by in- and outsiders becomes more and more difficult. This paper deals with a relatively new technique that appears to be suitable for solving these issues, i.e. anomaly detection based on profiling mobile users. Mobility pattern generation and behavior prediction are discussed in depth, before a new model of anomaly detection that is based on the Bayes decision rule is introduced. Applying this model to mobile user profiles proves the feasibility of our approach. Finally, a special emphasis is put on discussing privacy aspects of anomaly detection.

1. Introduction

Cellular radio networks gain more and more popularity and the amount of mobile communication will increase dramatically in the near future. Mobile users will no longer be restricted to the use of mobile phones. New network architectures like UMTS will place enhanced multimedia communication at the user's disposal. One major concern for the present and future cellular radio networks is security. But with increasing complexity of the networks the task of detecting, repulsing and preventing abuse by out- and insiders becomes more and more difficult. Obviously it is not possible to make any system absolutely secure with the currently known security techniques like e.g. authentication and encryption. This is related to the fact that a lot of attacks are simply based on software flaws and design errors, which may often be intelligently combined in order to open the door to any system under attack. One recent example is the cloning of GSM cards [4].

Additional countermeasures are therefore needed. One possible technique is anomaly detection based on the profiling of mobile users. Anomaly detection tries to detect the abnormal use of a system, i.e. a behavior which is significantly different from the usual behavior of a user. It is not restricted to any specific network environment. As a matter of fact, an anomaly detection component is a major building block within most available intrusion detection systems (IDS).

In this paper we focus on the application of anomaly detection techniques to mobile networks, an issue which has its own specific security requirements and user characteristics. We introduce a new model of anomaly detection which is based on the Bayes decision rule. A special emphasize is put on the privacy protecting aspects, and it is shown that, while a profiling function may seriously offend the privacy of one user, it can add value to another user.

Our paper is subdivided into seven sections. After this introduction we give a general outline of different anomaly detection models. Thereafter we introduce our own model based on the Bayes decision rule. In section 4 we apply this rule to generate a profile for a mobile user and prove its feasibility. Based on these results we continue with the description of general application scenarios and respective protocols. In this context we discuss privacy related issues. Afterwards we compare our approach with related work from other researchers and conclude with an outlook on future work.

2. Anomaly Detection

The first work in the field of anomaly detection has been done over a decade ago, focusing on main frame scenarios. With the rise of more complex data- and telecommunication networks like the Internet and mobile networks the designers of anomaly detection systems have to face new challenges resulting from the more distributed nature of these networks.

[23] lists three main statistical models currently used for anomaly detection:

- operational model

- mean and standard deviation model
- time series model

The operational model is based on thresholds, i.e. an alarm is raised if a variable observed (e.g. the number of login attempts) reaches a certain threshold. The mean and standard deviation model raises an alarm if an observation does not lie within a given confidence interval. The time series model takes the time at which an event takes place into account. If the probability for that event at that particular time is too low an alarm is raised.

Anomaly detection systems have major advantages compared to other intrusion detection approaches (see e.g. [23]) as they:

1. do not require any a priori knowledge of the target system, and
2. provide a way to detect unknown attacks.

But there also exist serious disadvantages which have to be considered before applying any of these techniques:

1. Not all users actually have a normal or standard behavior.
2. A user can slowly change his behavior over time from "good" to "bad", i.e. fool the system by slow long-term attacks.
3. The privacy of the users can be seriously injured.

We will come back to discuss these disadvantages and their relevance to mobile networks later. First of all we introduce our general approach towards anomaly detection, which is based on the Bayes rule.

3. The Bayes Decision Rule

Our approach towards anomaly detection is based on the Bayes decision rule and can therefore be classified as a statistical approach.

The Bayes decision rule is widely used in statistical pattern recognition [6]. A pattern recognition problem can be described in the following way: A set of objects can be divided into a number of classes. For each of these objects we measure a couple of observable characteristics and combine them to a vector. This observation vector will be different for each object, thus we can interpret this vector as a random variable X. To classify a new object, we have to learn the probability distribution of X for each class. If we know these distributions, we can calculate the probability that an object with observation vector x belongs to class c, namely $P(c|x)$. Therefore we can classify a new object in the following way:

1. Measure the observation vector for the object.
2. Calculate the class probabilities $P(c|x)$ for every class.
3. Choose the class with the highest probability as the object's class.

This decision rule is called the *Bayes decision rule for minimum error rate*. It can be shown that every other decision rule yields even higher error rates than the Bayes rule.

The so-called *a-posteriori* probability $P(c|x)$ can be expressed as:

$$P(c|x) = \frac{P(c)p(x|c)}{p(x)}$$

$p(x|c)$ is the *class conditional* probability density of observing a vector x, $P(c)$ is the *a-priori* probability for class c, and $p(x)$ is the probability density of observing a vector x. Because $p(x)$ is constant for every class c for a particular observation vector x, all we have to do is to learn the class conditional probability density $p(x|c)$ and the a-priori probability $P(c)$. $P(c)$ can be calculated as the relative frequency of observing a vector of class c. E.g. if we observe n vectors and n_1 vectors of them are of class c_1, the empirical probability $\hat{P}(c_1)$ can be calculated as

$$\hat{P}(c_1) = \frac{n_1}{n}$$

$p(x|c)$ is more difficult to learn. A simple technique is the use of histograms: We divide the vector space into intervals, count the number of vectors falling into every interval and then estimate the probability of vectors within this interval as the proportion of the number of vectors within this part compared to the number of all vectors. This technique only works if the number of intervals is small compared to the number of vectors (e.g. low dimension of the vector space).

4. Mobile User Profile Generation

We have applied the Bayes decision rule towards the generation of user profiles within GSM mobile networks.

4.1. The User and the Network

The network of GSM [12] is distributed in order to allow the reuse of transmission frequencies. Several *Mobile Switching Centers* (MSC) and local databases (*Visitor Location Register*, VLR) are distributed in the system serving their respective local areas (*MSC-areas*). The MSC-area in turn is subdivided into several *Location Areas* (LA), and an LA is subdivided into several cells, which are actually the smallest unit of the cellular network. Within a cell the mobile subscriber is able to call anyone and is reachable for

everyone. As the subscribers are free to go everywhere in the service area, it is obvious that they will enter and leave cells. Therefore, location information must be managed. Taking into account some performance considerations, the location information maintained in the network is in terms of LAs. Currently management of such data is organized using a central database, the *Home Location Register* (HLR). A centralized HLR stores data on subscribers and mobile stations. When a subscriber enters a new LA served by a new MSC, only the relevant data is downloaded to the VLR.

If a mobile terminated call from the public switched telephone network to a GSM subscriber has to be established, the call is routed to a gateway, called *Gateway MSC* (GMSC). GMSC interrogates with the called mobile user's HLR. The HLR requests the currently serving VLR to find out the routing number of the visited MSC. After receiving the mobile subscriber's number, the GMSC forwards the call to the terminating MSC. The MSC initiates the transmission of a paging message, i.e. the MSC pages the *Mobile Station* (MS) with a paging broadcast to all cells of the location area, as the exact cell of the user is unknown.

Location Update (LUP) and paging procedure are the basic operations for tracking and finding a user. Both procedures work with the same granularity of location information, e.g. five cells in an LA. Therefore mobile networks already provide the basic functionality which is necessary to track and profile users. With the use of more adaptive and dynamic tracking and paging algorithms, i.e. the collection and computation of more information about the users, it is possible to define individual mobility and behavior patterns. This additional knowledge about the users can be used to protect them and the network providers, because the availability of user profiles allows applying anomaly detection techniques.

Actually the first commercial versions of software which provide a certain degree of user protection through the GSM functionality itself have shown up on the market [3]. These systems physically hide a GSM terminal within a vehicle, which transmits a signal for localization purposes. The intended use is fleet management, vehicle tracking, vehicle recovery, rental services, emergency services and insurance.

Our approach goes one step further. We do not restrict ourselves to single positions a user takes during a trip, we also consider the different routes he normally takes over a longer period of time. This provides us with an elaborate user profile.

Take e.g. the following scenario into account: The mobile phone card of user A has been cloned by attacker X. If the GSM network is able to learn the routes user A normally takes in the sense that it can predict the time at which he is passing through single cells, the user can be tracked. In case of any anomaly, i.e. major deviations from the route, longer residence times or unusual cells, the user can be checked for its current status. Thereby misuse through the cloned phone card by attacker X can be detected by comparing the time and place of the call with the user's standard behavior.

One of the main questions is how the GSM network can learn the routes of the user, i.e. predict which is the most probable cell for a mobile station at a certain time, and what prediction level can be achieved. The next section proposes two algorithms for this learning process.

4.2. Decision Algorithms for Mobile Networks

The Bayesian Algorithm. If we interpret the time of day as the observation vector and the cells as classes, we can solve our problem of finding the most probable cell for a mobile station at a certain time of day with the Bayes rules as follows:

We divide the time axis into intervals (e.g. one second) with length δt. Using the movement patterns we generate a sequence of observation vectors $t_{1,c_1}, t_{2,c_2}, \ldots, t_{n,c_n}$ with $t_{i+1,c_k} - t_{i,c_l} = \delta t$. Using these vectors we estimate the distributions $P(c)$ and $p(x|c)$ by the respective empirical distributions $\hat{P}(c)$ and $\hat{p}(x|c)$ and calculate with the help of these distributions the most probable movement profile.

In the following the location algorithm using the Bayes decision rule will be called *Bayesian algorithm*.

Mean Residence Times. A simpler but efficient technique is the calculation of the mean residence times for each cell. Here we first have to find out the different movement patterns of a mobile station. Then we can use the residence times of the samples of each movement pattern to calculate the mean residence time for each cell. According to these times we can construct a movement profile.

We consider this simple algorithm, which we call the *Average algorithm*, as an example of a domain specific algorithm, while the Bayesian algorithm is universally applicable.

4.3. Evaluation Modeling

The predictability of the position of a mobile station depends on several factors such as:

- the number of movement patterns, and
- the variability of residence times in cells.

Even if the correct profile is chosen from the number of profiles (movement patterns), the residence times can degrade the performance of a prediction algorithm. Therefore two basic models have been developed for the performance evaluation of our approach:

- a model of the simulation surface and the routes of the mobile station, and

- a model of the residence times of the mobile station in the cells.

Surface and Route Model. For the surface model we have divided the surface into equally wide squares. Every square models one cell, and every location area contains the same number of cells. Every cell has got one unique number for identification. By varying the density of the cells different scenarios can be modeled (town, motorway). The routes which were used for the simulation were derived from the following real routes in Germany:

Scenario	Route	Distance	Mean Speed	Cell Size
town	Köln-Longerich Köln-Hochkirchen	19 km	40 km/h	1 km
motorway	Aachen-Europaplatz Köln-Longerich	74 km	100 km/h	3 km

In reality the base stations are located beside the roads to save signalling costs for LU and handover. Therefore we modified the routes on the simulation surface so that the mobile station crosses a cell only in vertical or horizontal direction. Thus a cell sequence can be found by the coordinates of the starting point and the endpoint and the sequence of points at which the movement direction changes.

Residence Times. The modeling of the residence times of a mobile station in the cells is more difficult. Instead of being fixed they vary around a certain mean. These variations can be caused by a lot of different factors like the actual traffic situation, the weather or the driver himself. Nonetheless, the driving time regarding a given route cannot fall below a minimal value because of the maximum speed of a vehicle and the speed limits.

To derive a model for the residence times, we need statistics of the behavior of individual users. Unfortunately such statistics do not exist. So we decided to use an investigation about driving time variations of bus routes [5]. In this investigation the driving times of the busses of three bus routes in Dortmund (Germany) have been measured. One main result was that the variations of the driving times between two bus stations can be described by an Erlang-k-distribution. The Erlang-k-distribution is zero in the origin, increases to its maximum and approaches zero again for infinity. For $k = 1$ the Erlang distribution is equivalent to the exponential distribution, for large k to the normal distribution. So, if we assume that the driving times of a car from cell border to cell border are comparable to the driving times of a bus from bus station to bus station, we can model the residence times for a mobile station in a cell by an Erlang-k-distribution.

This assumption is also backed by other research in this field [16].

We can motivate the use of an Erlang-k-distribution by the following rationale: As a participant mostly uses the same route through a cell, we can estimate a minimal residence time T_{min} by

$$T_{min} = \frac{S_{cell}}{V_{max}}$$

and the actual driving time T as

$$T = T_{min} + \Delta T$$

with S_{cell} being the length of the route through the cell, V_{max} the maximal speed of the user and ΔT the variation of the driving time which is restricted by $0 \leq \Delta T \leq \infty$. As a first approximation we can then model the time ΔT as an exponentially distributed random variable [2, 13, 17].

This simple model can be refined if we consider the fact that a route through a cell is not completely uniform, but consists of many small pieces with different driving conditions. So we can find a better approximation to reality if we model the run through a cell as a sequence of runs through smaller cells with i.i.d. exponentially distributed ΔT_{sub_i} (with rate λ, resp.). Put it in other words: The run through a cell may be inhomogeneous, i.e. the delay may vary for different parts of the cell. Therefore the cell is divided into subcells of such a respective size that identical delays are experienced while crossing each of these subcells. This may be modelled by a Poisson process X with arrival rate λ. Then the time between two arrivals in the Poisson process is equivalent to the time which is necessary for crossing one of the k subcells. Hence, the probability that crossing the whole cell takes longer than t equals the probability that it takes longer than t until k arrivals have taken place, i.e. the probability that there are less or equal $k-1$ arrivals between 0 and t [10]:

$$\begin{aligned} P(\Delta T > t) &= P(X \leq k-1) \\ &= \sum_{j=0}^{k-1} \frac{(\lambda t)^j}{j!} e^{-\lambda t} \\ F_{\Delta T}(t) &= P(\Delta T \leq t) \\ &= 1 - P(\Delta T > t) \\ &= 1 - \sum_{j=0}^{k-1} \frac{(\lambda t)^j}{j!} e^{-\lambda t} \end{aligned}$$

$F_{\Delta T}(t)$ is the distribution function of the trip time for the cell which is equal to the Erlang-k-distribution function.

4.4. Metrics

To evaluate the performance of the prediction algorithms we use the *Mean Prediction Level* as a metric.

The Mean Prediction Level (MPL) is the empirical probability that a mobile station is actually in the expected cell. This has to be verified by the verification function of the anomaly detection system, which compares the actual cell of the user with the expected cell(s). Let hit be the number of successful verifications and $miss$ the number of unsuccessful verifications. Then the MPL can be calculated as

$$MPL = \frac{hit}{hit + miss}$$

4.5. General Results

The following subsections present the MPL values and the 90% confidence intervals, which are obtained for different precision (different number of cells in an LA), for both prediction algorithms (Bayesian and Average algorithm) and the following verification strategies:

- Verification based on LA
  The most probable cell is taken from the profile and the corresponding LA is compared with the actual LA of the user.
- Verification based on an $2n$-minute interval
  The actual cell of the user is compared with the cells which fall into the time interval $(T - n, T + n)$ around the most probable cell.
- Verification based on n cells
  The actual cell of the user is compared with the n cells surrounding the most probable cell.

These different verification strategies define the sensitivity of the anomaly detection system concerning anomalous behavior. The first strategy accepts the highest degree of variance in the user's behavior, while the other two strategies are more stringent.

The results for each of the strategies are given in the following subsections.

Verification based on LA. Fig. 1 presents the MPL for the motorway scenario. We can notice the same evolution of the MPL for both algorithms. The stability of the system, i.e. a state for which new data do not improve the learning result, is reached after 15 days (in the following this state will be called *convergence state*). From this point the curve varies in a convergence interval of width 0.003, i.e. between 0.952 and 0.955. This also means that at least 15 dates must be collected to guarantee an efficient prediction and localization.

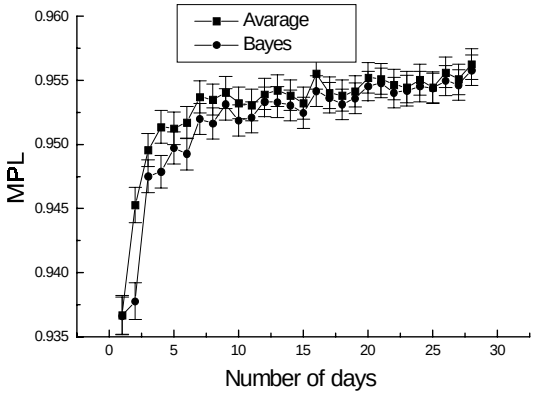

Figure 1. MPL - Verification based on LA for the motorway scenario.

We can notice that this strategy obtains a very good result, because the probability of a successful verification (for this type of user) reaches a value above 90% (0.936) at the second day.

Fig. 2 presents the result of this strategy for a city. At the convergence state, which is reached for both algorithms after 15 days, the MPL is 0.835. This value is 20% lower than the MPL value on the motorway. The reason for this is the small cell size in the city (1 km diameter) and the resulting small LA size.

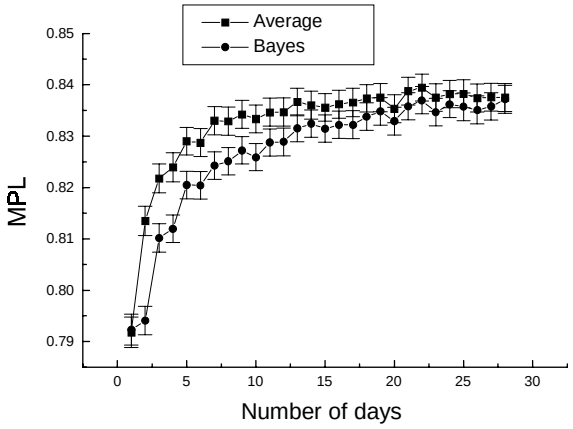

Figure 2. MPL - Verification based on LA for the city scenario.

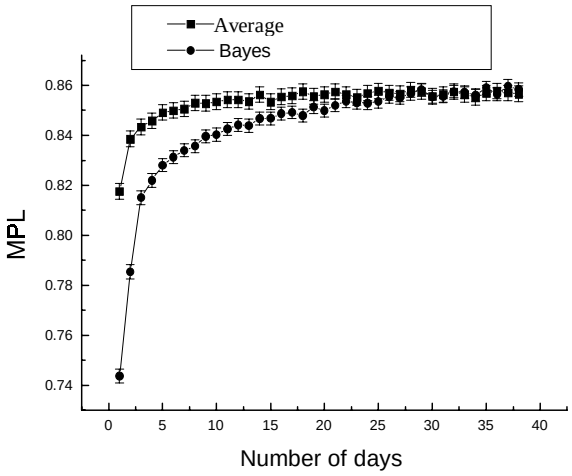

Figure 3. MPL - Verification based on 3 cells.

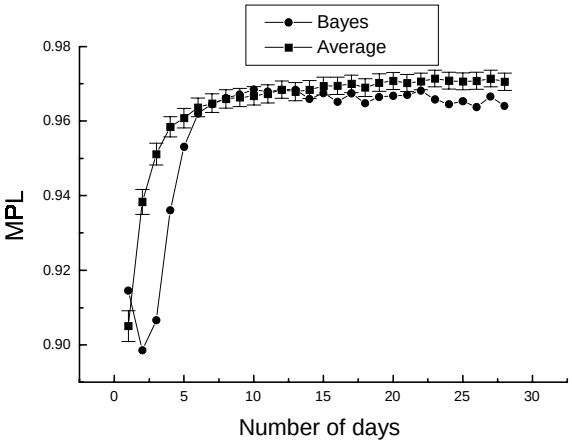

Figure 4. MPL - Verification based on a $2n$-minute interval (city).

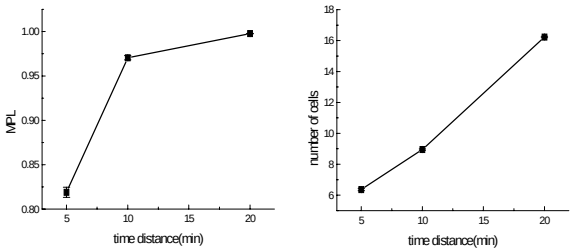

Figure 5. The effect of the time interval on the MPL (left) and the number of cells (right).

Verification based on 3 cells. Fig. 3 presents the results for the motorway scenario with a verification strategy based on three cells. The estimation of the three cells is performed for both algorithms in different ways. The Average algorithm estimates for the time of the incoming call (T) the most probable cell and its direct neighbors from the profile matrix. The Bayesian algorithm can estimate directly the three most probable cells. At the convergence state, which is reached with 22 days, the MPL varies around 0.85. For the Bayesian algorithm the first value is located around 0.74, so the gradient of the learning curve is 11% there. Considering the first value we can notice that the same value can be achieved if we try to locate the user in one cell. This is because, if we use only one day for estimating the profile, then only one most probable cell exists.

Verification based on an $2n$-minute interval. For the verification strategy we estimate the cells at time (T) which lie in the (T-n;T+n) interval. The investigation of this strategy is done for $n = 20$ minutes and for the city model. Both curves start with 0.91, after the 10th day they converge at 0.97 (see Fig. 4).

With this strategy we obtain nearly the same results as by the verification based on LA. In both cases the probability of a successful verification is above 95%. The advantage of this strategy is the possibility of a dynamic estimation of the LA size for each user. If a user is called seldomly we can choose a big time interval, for a user who is called often the time interval is small. The size of the interval determines the number of cells in a search area. The question is, how many cells should be included into an interval, if we consider e.g. a city with the cell size of 1 km. The answer is presented in Fig. 5 (right). On the abscissa the different time intervals ($n = 5, 10, 20$ minutes) are spread. The investigations are performed for the convergence state (based on 30 dates for estimating a profile). We can see that for $n = 5$ minutes the number of cells is about 6.5, for $n = 10$ minutes 9 and for $n = 20$ minutes 16. Fig. 5 (left) presents the dependence between time interval and MPL. For intervals longer than $n = 8$ minutes the user can be located with a probability of more than 95%.

The number of cells was determined not only for the driving time but for the whole observation period. In reality an observation period has a duration of one day. In the simulation the subscriber moves around half of the time, and the other time he stays at the same place. That means, if a user stays at the same place for a long time (e.g. at work), then the average number of cells converges to 1, because most of the calls are routed to the same cell. For the investigations

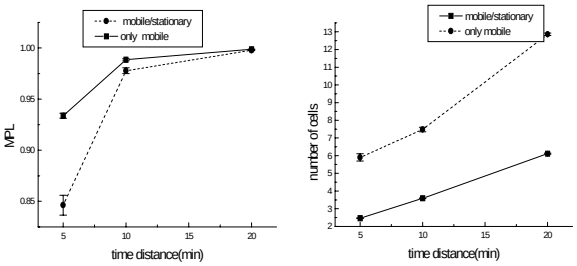

Figure 6. The effect of the length of the stationary state on the MPL (left) and the number of cells (right) in dependence of the time interval.

above we have assumed that the subscriber spends a part of the observation period at the same place. The next figure (Fig. 6) shows how the MPL rates and the number of cells (which were used for paging) change if the investigations are limited to the period where the subscriber moves.

These results refer to the motorway and are presented together with the results (mobile/stationary). The values for the five minute time interval are different, the values for the ten and twenty minute time intervals are approximately the same. The reason for this is that for big time intervals the user can be located anyway. For small time intervals it is more difficult to locate the user during the trip. This long duration of immobility influences the result.

Hence we need a classificator which indicates if a subscriber is only mobile or mobile and stationary to choose the appropriate time interval.

Notice that our current approach does not include a dynamic learning process in the sense that it considers already verified actual user positions. In that sense our simulation is a worst-case analysis and the result can be improved by introducing the described level of dynamic learning.

5. Application Scenarios and the Privacy of the Users

The results described above prove that it is possible to profile certain kinds of mobile user with a high quality concerning precision, speed and stability. Two general questions arise:

- Which kind of users can actually be profiled?
- Which of these users would like to be profiled?

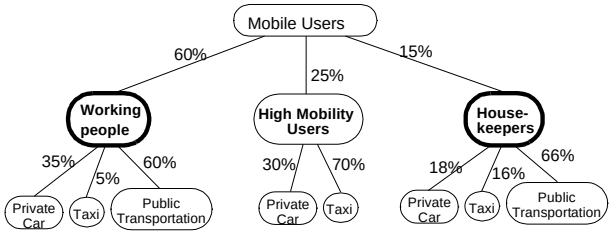

Figure 7. Categorization of mobile users according to their mobility behavior.

5.1. UMTS User Categories

The answer to the first question leads us to a basic observation, which actually provides the main base for our work. The UMTS RACE specification dealing with mobility management explicitly takes the different mobility behavior of the users into account. For this reason the model considers different user classes, environments, geographical areas like e.g. workers, driving in a bus from the rural to the metropolitan area[1]. The classification is done according to the statistics taken from different conurbations: London (UK), Randstad (Netherlands) etc. A categorization of mobile users considering their mobility behavior is shown in Fig. 7 [11][2].

In a further investigation UMTS discusses the different aspects of user mobility. There the mobility is defined as a movement from one geographical location to another. The categorization of the roaming area is made according to the frequency the user visits these areas. The area with the highest frequency is the daily used access domain. Within this domain the user roams daily, for instance the trip of a worker from his home to his working place. Statistics about the starting point (generation model), the endpoint (attraction model), movement duration, time of day or day of week versus percentage of user types (e.g. car driver) can be found in the RACE specification.

The general conclusion for our work must be that the classification shows that approximately 75% of the users can be considered as potential candidates for the successful application of anomaly detection techniques. Only the high mobility users, for which it must be expected that they establish much more chaotic behavior patterns, are not directly suited for our approach. The other user groups can be expected to behave according to more stringent behavior patterns, simply because of their social environment.

[1] RACE considers the following areas starting from the center of the city: metropolitan, urban, suburban and rural.

[2] To avoid any conflict with ETSI, we took a publication that follows the work of RACE and was partly written from the same authors of the RACE document.

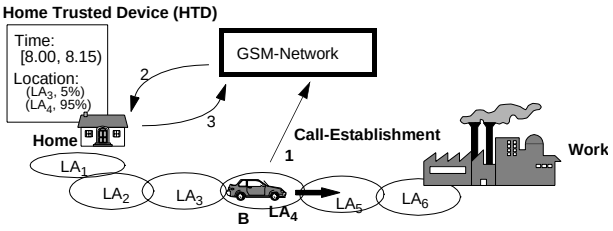

Figure 8. Trustworthy maintenance of mobility profile in a HTD.

5.2. Profiling and Privacy

While this classification and the related numbers answer the question what kind of users can actually be profiled, the second question remains to be answered: Which of these users would like to be profiled? For the mobile network scenario depicted above the profiling adds an additional level of security to specific groups of users. Other users will not be too enthusiastic about the possibility of profiling their behavior. In order to protect this sensitive information the control over the profiling and anomaly detection algorithms must obviously be located at a unit that the user trusts and/or controls. No potential communication partner or third party (including the network operator) should be able to have access to the users profile.

An approach to store profiles in conformity with these security requirements is to keep them in a trustworthy environment. [14], [20] and [21] propose to store confidential information in a personal digital assistant or "personal communication bodyguard". A *Home Trusted Device* (HTD) is the most common realization of such a trusted and private environment. In the following we explain the role it can play in profile management and anomaly detection in a mobile network (see Fig. 8).

The above scheme of a learning, predicting and controlling HTD, which supplements the classical GSM network, can be realized as a modular extension. User B repeats his daily movement from home to work and back. The MS continues to use the classical GSM location tracking algorithm and collects the daily repeated moving data. This information is transmitted to the HTD every day in a secure way by connecting the MS directly with HTD_B. Secret key and authentication procedures guarantee the integrity of the data. After this initial phase the HTD generates a user specific profile. The profile is stored in the HTD.

If a call has to be established (1), the GSM location tracking algorithm sends its current location information to the HTD (2). The HTD compares the location of the user with his stored profile. If the location information given by the GSM net corresponds to the stored profile, i.e. the verification results in a positive answer, the call can be established. Otherwise special measures, which we leave unspecified here, have to be taken. The HTD sends its answer back to the GSM net[3] (3). This request/response signalling should be protected using a symmetric cryptographic system.

The measure taken in case of a deviation from the stored profile and the sensitivity of the deviation must also be controlled by the user. In general the mobile network scenario is not that sensitive to certain attacks like other networks. Recalling the argument that a general anomaly detection systems can be fooled by a user, which slowly changes his behavior over time from "good" to "bad", does not apply to our scenario, because we do not have to deal with these kinds of internal attacks. In case of a slow shift of the behavior pattern the profile must be updated, because it can be considered to be a valid change of the user's behavior. An attacker must either steal a mobile device or clone the card. In the first case only one user with a specific ID actually uses the network, in the second case two users are present in the net. Both attacks do not give the attacker the time to slowly modify the stored profiles.

6. Profiling and Anomaly Detection - Related Work

Work related and relevant to our own approach comes from two fields:

- Network Oriented Location Management, and
- Anomaly Detection.

Concerning the network oriented location management several investigations make the assumption that the mobility of the users can be foretold [15], [18], [22].

The network in [15] observes the mobility of every user and generates a profile for an individual user with the following content: for each period $[t_i, t_j)$ the network handles a set of location area identifiers $a_1, ..., a_n$ combined with $\alpha_1, ..., \alpha_n$. α_i is the probability that the user is in the respective location area a_i. Having this information in the network "memory", the network does not need to explicitly track the user's mobility if the user follows the known mobility pattern. If a call has to be established, the appropriate set of a_i's and α_i's will be chosen considering the calling time. The network will then broadcast in the different a_i's according to their probabilities. Having this profile a first check will be the calculation of the predictability level of the user. Therefore Tabbane defines in [22] the mobility predictability level as $MPL := \sum_{i=1}^{N} \frac{\alpha_i}{i}$. This parameter is a measure of how predictable the pattern of the users

[3] This check can be performed in parallel to the connection setup in order to accelerate the call establishment.

are. In the subsequent work he evaluates his prediction by using the MPL as an input parameter. He concludes that using a mobility profile results in a "great amount of resource (radio as network)" savings especially for users with high MPL values.

Our approach follows these dynamic location management concepts (especially the work of [22]) and we have also shown that this approach, like the one mentioned above, can be used to reduce the location update and paging costs [9]. Nonetheless the focus in these works has been put on cost aspects and not on security aspects. To use these profiles in the sense of an anomaly detection system for a mobile network has to our knowledge not been considered before.

The second branch of relevant work is the general research in the field of anomaly detection, especially as a building block of a broader IDS. Examples are the IDES Statistical Anomaly Detector [7], which monitors a computer system and adaptively learns what is normal for individual users and groups of users. Based on these data it identifies potential intrusions. For the profiling the so-called multivariate method is used (see also [8]). Statistical profiles and their generation can also be found in many other systems and publications (see e.g. [1] [19]). However, to our knowledge, no work exists so far which bases its learning mechanism on the Bayes decision rule. In addition the currently known anomaly detection mechanisms have not been applied to the mobile network scenario depicted above.

7. Conclusions and Outlook

In this paper we have proposed a new algorithm for the profiling of mobile users, which is based on the Bayes decision algorithm. It has been shown that this approach can successfully be applied in order to provide advanced security features for mobile network users. One of our main concerns has been the discussion of privacy problems related to the profiling of users.

Our future work will concentrate on the adaptation of our profiling technique to standard data communication networks like the Internet. We will investigate the performance of the Bayesian algorithm for Intranet and LAN scenarios, the classical application fields of intrusion detection systems.

Further investigations will extend our approach to provide additional security to a mobile agent environments, as obviously there exists a straightforward mapping between the mobile network scenario (users moving from cell to cell) and an mobile agent scenario (agents moving from server to server).

References

[1] D. Anderson, T. Lunt, H. Javitz, A. Tamaru, and A. Valdes. Safeguard final report: Detecting unusual program behavior using the NIDES statistical component. Technical report, Computer Science Laboratory, SRI International, Menlo Park, CA, December 1993.

[2] P. Camarda, G. Schiraldi, and F. Talucci. Mobility modeling in cellular communication networks. In *21st IEEE Conference on Local Computer Networks*, Minneapolis, Minnesota, October 1996.

[3] Cellpoint Systems AB, http://www.cellpt.com/. *GSM Positioning*.

[4] Computer Science Division, University of California, Berkeley, http://www.isaac.cs.berkeley.edu/isaac/gsm-faq.html. *Internet Security, Applications, Authentication and Cryptography (ISAAC) Research Group*.

[5] F. Dauber. *Einflußgrößen auf die Bedienqualität von Buslinienverkehren unter besonderer Beachtung von Fahrplan- und Folgezeitenabweichungen*. PhD thesis, RWTH Aachen, Germany, 1986.

[6] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. J. Wiley, New York, 1973.

[7] H. Javitz and A. Valdes. The SRI IDES statistical anomaly detector. In *Proceedings of the IEEE Symposium on Research in Security and Privacy*, pages 316–326, May 1991.

[8] H. Javitz and A. Valdes. The NIDES statistical component description and justification. Technical report, Computer Science Laboratory, SRI International, Menlo Park, CA, March 1994.

[9] D. Kesdogan, M. Zywiecki, and K. Beulen. Mobile user profile generation - a challenge between performance and security. In *Proceedings of the 2nd Workshop on Personal Wireless Communications (Wireless Local Access)*, Frankfurt am Main (Germany), December 1996.

[10] H. J. Larson. *Introduction to the Probability Theory and Statistical Interference*. Wiley Series in Probability and Mathematical Statistics. J. Wiley & Sons, New York, 1982.

[11] J. Markoulidakis, G. Lyberopoulos, D. Tsirkas, and E. Sykas. Evaluation of location area planning scenarios in future mobile telecommunication systems. In *Wireless Networks 1*, 1995.

[12] M. Mouly and M.-B. Pautet. The GSM system for mobile communications. ISBN 2-9507190-0-7, 1992.

[13] S. Nanda. Teletraffic models for urban and suburban microcells: Cell sizes and handoff rates. In *Wireless Communications, Future Directions*, Dordrecht, Boston, 1993.

[14] A. Pfitzmann. Technischer Datenschutz in öffentlichen Funknetzen. *Datenschutz und Datensicherheit (DuD)*, pages 451–463, August 1993.

[15] G. Pollini and S. Tabbane. The intelligent network signalling and switching costs of an alternate location strategy using memory. In *IEEE 43th VTC*, 1993.

[16] S. Rappaport. Blocking, hand-off and traffic performance for cellular communication systems with mixed platforms. In *IEEE Proceedings-I, Vol. 140, No. 5*, October 1993.

[17] S. Rappaport and L. Hu. Microcellular communication systems with hierarchical macrocell overlays: Traffic performance models and analysis. In *Proceedings of the IEEE, Vol. 82, No. 9*, September 1994.

[18] C. Rose and R. Yates. Ensemble polling strategies for mobile communication networks. In *IEEE 46th VTC*, 1996.

[19] M. Sebring, E. Shellhouse, M. Hanna, and R. Whitehurst. Expert systems in intrusion detection: A case study. In *Proc. of the 11th National Computer Security Conf.*, Baltimore, MD, October 1988.

[20] M. Spreitzer and M. Theimer. Scalable, secure, mobile computing with location information. *Communications of the ACM*, 36(7), 1993.

[21] M. Spreitzer and M. Theimer. Architectural considerations for scalable, secure, mobile computing with location information. In *Proceedings of the 14th International Conference on Distributed Systems*. IEEE, 1994.

[22] S. Tabbane. An alternative strategy for location tracking. *IEEE Journal on Selected Areas in Communications*, 13(5), June 1995.

[23] A. B. Tucker Jr., editor. *CRC Computer Science and Engineering Handbook*. CRC Press, December 1996.