FRAUD DETECTION AND MANAGEMENT IN MOBILE TELECOMMUNICATIONS NETWORKS

P Burge, J Shawe-Taylor, C Cooke, Y Moreau, B Preneel, C Stoermann.

Royal Holloway University of London, England; Vodafone, England; ESAT.K.U.Leuven, Belgium; Siemens A.G. Germany.

ABSTRACT: *This paper discusses the status of research on detection of fraud undertaken as part of the European Commission-funded ACTS ASPeCT (Advanced Security for Personal Communications Technologies) project. A first task has been the identification of possible fraud scenarios and of typical fraud indicators which can be mapped to data in Toll Tickets. Currently, the project is exploring the detection of fraudulent behaviour based on a combination of absolute and differential usage. Three approaches are being investigated: a rule-based approach and two approaches based on neural networks, where both supervised and unsupervised learning are considered. Special attention is being paid to the feasibility of the implementations.*

1. INTRODUCTION

It is estimated that the mobile communications industry loses several million ECUs per year due to fraud. Therefore, prevention and early detection of fraudulent activity is an important goal for network operators. It is clear that the additional security measures taken in GSM and in the future UMTS (Universal Mobile Telecommunications System) make these networks less vulnerable to fraud than the analogue networks. Nevertheless, certain types of commercial fraud are very hard to preclude by technical means. It is also anticipated that the introduction of new services can lead to the development of new ways to defraud the system. The use of sophisticated fraud detection techniques can assist in early detection of commercial frauds, and will also reduce the effectiveness of technical frauds.

One of the tasks of the European Commission-funded ACTS project ASPeCT (Advanced Security for Personal Communications Technologies) ASPeCT (1) is the development of new techniques and concepts for the detection of fraud in mobile telecommunication networks. This paper intends to report on the progress made during the first year. For a more detailed status report, the reader is referred to ASPeCT (2).

The remainder of this paper is organised as follows: Section 2 discusses the identification of possible fraud scenarios and of fraud indicators; Section 3 discusses the general approach of user profiling; Sections 4 and 5 present respectively the rule-based approach and the neural network based approach to fraud detection.

2. POSSIBLE FRAUDS AND THEIR INDICATORS

The first stage of the work consists of the identification of possible fraud scenarios in telecommunications networks and particularly in mobile telephone networks. These scenarios have been classified by the technical manner in which they are committed; also an investigation has been undertaken to identify which parts of the mobile telecommunications network are abused in order to commit any particular fraud. Other characteristics that have been studied are whether frauds are technical fraud operated for financial gain, or they are fraud related to personal use - hence not employed for profiteering. A further classification is achieved by considering whether the network abuse is the result of *administrative fraud*, *procurement fraud*, or *application fraud*.

Subsequently, typical indicators have been identified which may be used for the purposes of detecting fraud committed using mobile telephones. In order to provide an indication of the likely ability of particular indicators to identify a specific fraud, these indicators have been classified both by their *type* and by their *use*.

The different types are :-

- **usage indicators**, related to the way in which a mobile telephone is used;
- **mobility indicators**, related to the mobility of the telephone;
- **deductive indicators**, which arise as a by-product of fraudulent behaviour (e.g., overlapping calls and velocity checks).

Indicators have also been classified by use:-

- **primary indicators** can, in principle, be employed in isolation to detect fraud;
- **secondary indicators** provide useful information in isolation (but are not sufficient by themselves);
- **tertiary indicators** provide supporting information when combined with other indicators.

A selection has been made of those scenarios which cannot be easily detected using existing tools, but which could be identified using more sophisticated approaches.

The potential fraud indicators have been mapped to network data required to measure them. The information required to monitor the use of the communications network is contained in the Toll Tickets.

Toll Tickets are data records containing details pertaining to every mobile telephone call attempt. Toll Tickets are transmitted to the network operator by the switch which handles the originating or forwarded part of the mobile telephone call. They are used to determine the charge to the subscriber, but they also provide information about customer usage and thus facilitate the detection of any possible fraudulent use. An investigation has been completed of which fields in the GSM Toll Tickets can be used as indicators for likely fraudulent behaviour.

Before use in the fraud detection engine, the Toll Tickets are pre-processed. An essential component of this process is the encryption of all personal information in the Toll Tickets (such as MSISDN numbers). This allows for the protection of the privacy of users during the development of the fraud detection tools, while at the same time the network operators will be able to obtain the identity of suspected fraudulent users.

3. USER PROFILING

3.1 Absolute or differential analysis

Existing fraud detection systems tend to interrogate sequences of Toll Tickets comparing a function of the various fields with fixed criteria known as *triggers*. A trigger, if activated, raises an alert status which cumulatively would lead to an investigation by the network operator. Such fixed trigger systems perform what is known as an *absolute* analysis of the Toll Tickets and are good at detecting the extremes of fraudulent activity.

Another approach to the problem is to perform a *differential* analysis. Here we monitor behavioural patterns associated with the mobile telephone comparing its most recent activities with a history of its usage. Criteria can then be derived to be used as triggers that are activated when usage patterns of the mobile telephone change significantly over a short period of time; a change in the behaviour pattern associated with a mobile telephone is a common characteristic in nearly all fraud scenarios.

There are many advantages to performing a differential analysis through profiling the behaviour of a user. Firstly, certain behavioural patterns may be considered anomalous for one type of user, and hence potentially indicative of fraud, but which may be considered acceptable for another user. With a differential analysis flexible criteria can be developed that detect any change in usage based on a detailed history profile of user behaviour. This takes fraud detection down to the personal level comparing like with like enabling detection of less obvious frauds that may only be noticed at the personal usage level. An absolute usage system would not detect fraud at this level. In addition, however, because a typical user is not a fraudster, the majority of criteria that would have triggered an alarm in an absolute usage system will be seen as a large change in behaviour in a differential usage system. In this way a differential analysis can be seen as incorporating the absolute approach.

3.2 The differential approach

Most fraud indicators do not become apparent from an individual Toll Ticket. With the possible exception of an overlapping call or a velocity trap, we can only gain confidence in detecting a real fraud through investigating a fairly long sequence of Toll Tickets. This is particularly the case when considering more subtle changes in a user’s behaviour by performing a differential analysis.

A differential usage system requires information concerning the user’s history of behaviour plus a more recent sample of the mobile phone’s activities. An initial approach might be to extract and encode information from Toll Tickets and to store it in record format. This would require two windows or spans over the sequence of transactions for each user. The shorter sequence might be called the Current User Profile (CUP) and the longer sequence, the User Profile History (UPH). Both profiles could be treated and maintained as finite-length queues. When a new Toll Ticket arrives for a given user, the oldest entry from the UPH would be discarded and the oldest entry from the CUP would move to the back of the UPH queue. The new record encoded from the incoming Toll Ticket would then join the back of the CUP queue.

Clearly it is not optimal to search and retrieve historical information concerning a user’s activities prior to each calculation, on receipt of a new Toll Ticket. A more suitable approach is to compute a single cumulative CUP and UPH for each user, from incoming Toll Tickets which can be stored as individual records, possibly in a database. So that we maintain the concept of having two different spans over the Toll Tickets without retaining a database record for each Toll Ticket, we will need to decay both profiles before the influence of a new Toll Ticket can be taken into consideration. A straightforward decay factor may not be suitable as this will potentially dilute information relating to encoded parameters stored in the user’s profile. An important concern here is the potential creation of false behaviour patterns. Several decaying systems are currently being investigated.

3.3 Relevant Toll Ticket data

There are two important requirements for user profiling. At first, efficiency is of the foremost concern for storing the user data and for performing updates. Secondly, user profiles have to realise a precise description of user behaviour to facilitate reliable fraud detection. All the information that a fraud detection tool will need to handle is derived from the Toll Tickets provided by the network operator.

The following Toll Ticket components have been observed as being the most fraud relevant measures:

- Charged_IMSI (identifies the user)
- First_Cell_ID (location characteristic for mobile originating calls)
- Chargeable_Duration (base for all cost estimations)
- B_Type_of_Number (for distinguishing between National / International calls)
- Non_Charged_Party (the number dialled)
- Charging_Start_Time (call start time)

These components will continually be picked out of the Toll Tickets and incorporated into the user profiles in a cumulative manner.

It is also anticipated that the analysis of cell congestion can provide useful ancillary information.

4. RULE-BASED APPROACH TO FRAUD DETECTION

In ASPeCT, several approaches are taken to identify fraudulent behaviour. In the rule-based approach, both the absolute and differential usage are verified against certain rules. This approach works best with user profiles containing explicit information, where fraud criteria given as rules can be referred to. User profiles are maintained for the directory number of the calling party (A-number), for the directory number of the called party (B-number) and also for the cells used to make/receive the calls. A-number profiles represent user behaviour and are useful for the detection of most types of fraud, while B-number profiles point to *hot destinations* and thus allow the detection of frauds based upon call forwarding. All deviations from normal user behaviour resulting from the different analysing processes are collected and alarms will finally be raised if the results in combination fulfil given alarm criteria.

The implementation of this solution is based on an existing rule-based tool for audit trail analysis PDAT (Protocol Data Analysis Tool) Katzer and Mehlhart (3). PDAT is a rule based tool for intrusion detection developed by Siemens ZFE (Corporate Research and Development). PDAT works in heterogeneous environments, has the possibility of on-line analysis, and provides a performance of about 200 KB input per second. Important goals were flexibility and broad applicability, including the analysis of general protocol data, which is achieved by the special language PDAL (Protocol Data Analysis Language). PDAL allows the programming of analysis criteria as well as a GUI-aided (Graphical User Interface) configuration of the analysis at run-time.

Intrusion detection and mobile fraud detection are quite similar problem fields; the flexibility and broad applicability of PDAT are promising for using this tool for mobile fraud detection. The main difference between intrusion detection and mobile fraud detection seems to be the kind of input data. The recording for intrusion detection produces 50 MB per day per user, but only for the few users of one UNIX-system. In comparison, mobile telephone fraud detection has to deal with a huge amount of subscribers (roughly 1 Million) each of whom, however, produces only about 300 bytes of data per day. PDAT was able to keep all interim results in main memory, since only a few users had to be dealt with. For fraud detection, however, intermediate data has of course to be stored on hard disc. The main tasks were the introduction of user profiles stored in a data base and the realisation of a new protocol that allows PDAT to understand both user profile as well as Toll Ticket formats. Once established, PDAT provides a comprehensive infrastructure based on a GUI for showing alarms and for editing alarm criteria during runtime. The new architecture is depicted in Figure 1.

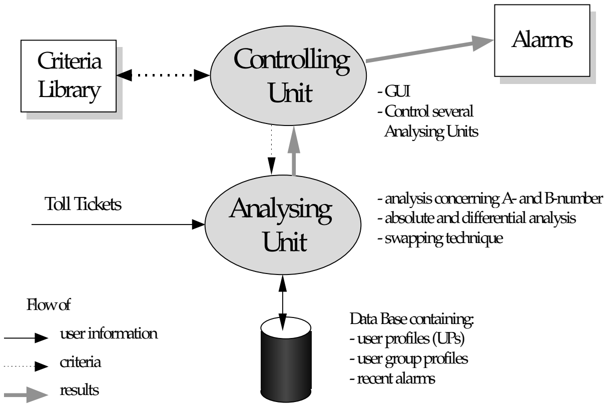

Figure 1 Architecture of rule-based fraud detection tool

5. NEURAL NETWORK BASED APPROACH TO FRAUD DETECTION

A second approach to identify fraudulent behaviour uses neural networks. The multiplicity and heterogeneity of the fraud scenarios require the use of intelligent detection systems. The fraud detection engine has to be flexible enough to cope with the diversity of fraud. It should also be adaptive in order to face new fraud scenarios, since fraudsters are likely to develop new forms of fraud once older attacks become impractical. Further, fraud appears in the billing system as abnormal usage patterns in the Toll Ticket records of one or more users. The function of the fraud detection engine is to recognise such patterns and produce the necessary alarms. High flexibility and adaptivity for a pattern recognition problem directly point to neural networks as a potential solution. Neural networks are systems of elementary decision units that can be adapted by training in order to recognise and classify arbitrary patterns. The interaction of a high number of elementary units makes it possible to learn arbitrarily complex tasks. For fraud detection in telephone networks, neural network engines are currently being developed world-wide (4,5). As a closely related application, neural networks are now routinely used for the detection of credit card fraud.

There are two main forms of learning in neural networks: unsupervised learning and supervised learning. In supervised learning, the patterns have to be *a priori* labelled as belonging to some class. During learning, the network tries to adapt its units so that it produces the correct label at its output for each training pattern. Once training is finished the units are frozen, and when a new pattern is presented, it is classified according to the output produced by the network. In unsupervised learning the system is allowed to find patterns or clusters in the data in the hope that these clusters will be useful or meaningful in some way, either directly or indirectly.

5.1 Supervised Learning.

For supervised learning we utilise a multi-layer perceptron. It is defined as follows. The network is composed of elementary units called neurons. Each neuron produces at its output a simple non-linear transformation of its inputs depending on the value of the weights of the network, thus:-

$$y = \sigma(\sum_{i=1}^{n} w_i x_i + w_0), \quad \text{where } \sigma(z) = \tanh(z) \text{ or}$$

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

Where x_i is the value on the i-th input line and w_i is the weight on that line.

The neurons are then arranged in a two-hidden-layer network with D inputs, H_1 hidden neurons in the first layer, H_2 hidden neurons in the second layer, and C outputs. The outputs z_m of the network can then be defined

as

$$h_{1_k} = \sigma(\sum_{l=1}^{D} w_{kl} x_l + w_{k0}), h_{2_l} = \sigma(\sum_{k=1}^{H_1} v_{ik} h_{1_k} + v_{i0}),$$

$$z_m = \sigma(\sum_{l=1}^{H_2} u_{lm} h_{2_l} + u_{l0})$$

.

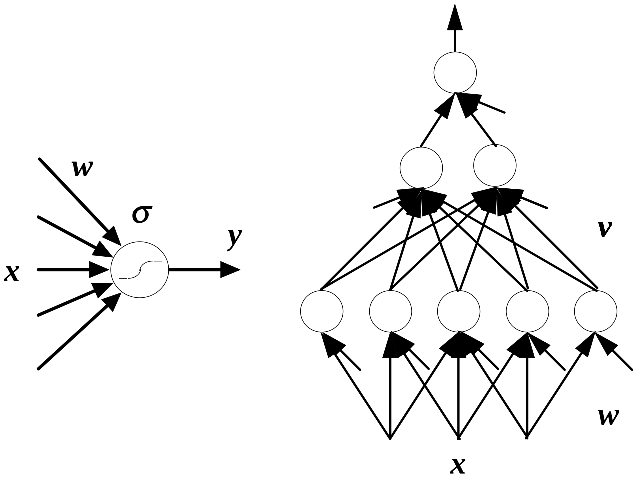

Figure 2 Sigmoidal neuron and multi-layer perceptron architecture

The main property of multi-layer perceptrons is that they can approximate any function of the input to an arbitrary degree of accuracy, provided that enough hidden neurons are available. They can achieve this approximation with a relatively small number of parameters.

For supervised learning, we organise the data available for design in a data set of labelled pairs D = $\{(X_1,Y_1),...,(X_K,Y_K)\}$, where Y_k is the fraud label (Y_k = 0 for normal behaviour, Y_k = 1 for fraud) associated to the *k*th pattern with features X_k extracted from the user profile. The training data consists for the first part of the calls made by 300 users from a two-month download of sanitised GSM Toll Tickets from the Vodafone network; the user's behaviour in this set is considered normal. It also consists for a second part of the calls made by 300 fraudulent users. For all 600 users, all the available Toll Tickets are processed through the front-end of the system to produce sequences of user profiles. We label the sequences of user profiles for the normal users as non-fraudulent. For the fraudsters, we study the evolution of the profiles over time to determine the beginning of the fraudulent behaviour; and we label the profiles as fraudulent during the fraudulent behaviour and as non-fraudulent otherwise.

The first step in training is to choose the architecture of the neural network, that is the number of layers and the number of neurons in each layer. Once we have chosen the architecture, the output of the network is a function of its input X_k and of the parameters w (the weights) of the neural network. There is a discrepancy between the output of the classifier $z(X_k,w)$ and the desired output Y_k. The training of the classifier consists of adapting the weights so as to minimise this discrepancy. The measure of discrepancy is quadratic, where w is found such that E is a minimum, thus:-

$$\mathrm{E} = \sum_{k=1}^{K} \left\| \mathrm{Y}_k - \mathrm{z}(\mathrm{X}_k, \mathrm{w}) \right\|^2$$

We achieve this minimisation using a gradient-descent method, namely the Levenberg-Marquardt algorithm. This powerful method is based on algebraic procedures, which permits, given a value of the weights, the determination of the necessary modification of the weights to result in an optimal reduction of the error. After we have modified the weights, we have a smaller error and a new value of the weights. A new correction to the weights is evaluated, so as to reduce the error as rapidly as possible. We repeat this procedure until the error stops decreasing. In order to maximise the performance on previously unseen data, we use the following procedure called cross-validation. The weights are adapted by minimising the error on the training set, but we observe the error on the validation set during this process; and we stop the minimisation when the error on the validation set reaches a minimum. We then estimate the expected performance on new data by computing the error on the test set.

We determine the optimal weights using the error minimisation procedure, but we have to repeat the procedure to search for a global optimum of the optimisation procedure, since gradient-descent methods are only guaranteed to converge to a local optimum. Furthermore, we have to repeat this procedures for different architectures of the neural network to determine the optimal one. Once we have found the optimal neural network, we simply have to use it on top of the front-end and it will produce an alarm value between 0 and 1 each time a Toll Ticket is presented to the fraud detection tool. This alarm is passed on to the operator.

5.2 Unsupervised Learning

With unsupervised learning, statistical user profiles are generated by the classification of GSM Toll Tickets into one of a set of Toll Ticket prototypes. The user profiles are comprised of a vector of values essentially counting the number of times a Toll Ticket prototype gets excited by the presentation of a Toll Ticket. The task of the system, after developing user profiles over a specified period, is to raise alarms when it is presented with profiles where the difference between the CUP and the UPH is outside the realms of normal usage. An alert status will be raised if the profiles are significantly different. Note that the system requires only clean (non fraudulent) data for training. This system therefore has the potential to detect new types of fraud as and when they occur.

Prototyping is a method of forming an optimal discrete representation of a naturally continuous random variable. The processing of continuous random variables by discrete systems generally reduces empirical information. Neural Networks are capable of forming optimal discrete representations of continuous random variables through their ability to converge, by lateral interaction, to stable uniformly distributed states.

The mapping of a continuous random variable X into a set of K discrete prototypes Q reduces the empirical information by the least amount if a uniform distribution {$\mathbf{P}(q_i) = \frac{1}{K}, i = 1 \ldots K$}, corresponding to the absolute maximum ($S_Q = \log K$) of information entropy, is assigned to Q.

Grabec (6) provides a way to extend this principal to multiple dimensions. When considering the set of all possible Toll Tickets, we clearly have a dimension to represent every parameter that we wish to include in the analysis. Each parameter in a Toll Ticket can assume a range of values and is thus itself a random variable. Grabec’s technique will allow us to create a number of prototypes that dynamically and uniformly span the set of samples taken from the space of possible Toll Tickets.

For a differential analysis we need to maintain the concept of two different spans over the Toll Tickets. We maintain the two profiles as probability distributions using two different decay factors α and β. When a Toll Ticket is presented to the system to update a user’s CUP, each element of the CUP is multiplied by decay factor α. The entry in the profile corresponding to the prototype i, that was excited by the presentation of the incoming Toll Ticket, is then incremented by an amount $(1-\alpha)$. Updating the CUP in this manner will maintain the profile as a probability distribution. After updating the CUP, a differential analysis is performed by the fraud engine on the CUP and UPH. Following presentation to the fraud engine, the UPH is updated using

$$H_i = \beta H_i + (1-\beta) C_i$$

where H_i and C_i represent the *i*th element of the UPH and CUP respectively. By applying a multiplicative decay factor, any counter in the profile corresponding to a prototype, once excited, will never actually decay to zero. To perform the differential analysis on the user profiles we use a measure known as the Hellinger distance.

$$d = \sum_{l=0}^{K} \left(\sqrt{C_i} - \sqrt{H_i} \right)^2$$

In detection mode the fraud engine again calculates d and if the resultant value is greater than a preset threshold, then an alert status is raised proportional to $\left| d - d_{threshold} \right|$. It is anticipated that these alert statuses will be prioritized for investigation.

Each of the techniques presented has its merits and also its drawbacks. Our current prototype enables all three systems to work on the same data within a common framework. One future system that we are investigating is the possibility of using the unsupervised learning neural network system to filter out very normal behavior so that the processing load is reduced for the rule-based system and neural network-based system that uses supervised learning.

References

1. ACTS AC095, project ASPeCT: February 1996 *Initial report on security requirements*. AC095/ATEA/W21/DS/P/02/B.

2. ACTS AC095, project ASPeCT: September 1996 *Definition of fraud detection concepts*. AC095/KUL/W22/DS/P/06/A.

3. Katzer, T. Mehlhart, C. Wolff: 1993. *PDAT - ein Protokolldaten-Analysewerkzeug fuer sichere Betriebssysteme und Anwendungen*. Unix in Deutschland - GUUG, Network Verlag, Hagenburg, Germany.

4. Barson, S. Field, N. Davey, G. McAskie, R. Frank: 1996 *The Detection of Fraud in Mobile Phone Networks.* Neural Network World, Vol.6, No. 4, pp. 477-484.

5. Yuhas 1993: *Toll-Fraud Detection*. Proceedings of the International Workshop on Applications of Neural Networks to Telecommunications, ed. J. Alspector, R. Goodman, and T.X. Brown, pp. 239-244, Lawrence Erlbaum Associates.

6. Grabec I, 1989 Self-Organisation of Neurons Described by the Second Maximal Entropy Principle, *Proceedings 1st IEE International Conference on Artificial Neural Networks*, London, Conference Publication NO 313, pp. 12-16.