Project Number	AC095
Project Title	ASPeCT: Advanced Security for Personal Communications Technologies
Deliverable Type	Intermediate
Security Class	Public

Deliverable Number	D13
Title of Deliverable	Fraud management tool : evaluation report
Nature of the Deliverable	Report
Document reference	AC095/SAG/W22/DS/P/13/2
Contributing WPs	WP2.2
Contractual Date of Delivery	May 1997 (Y03/M03)
Actual Date of Delivery	28 October 1997
Editor	Christof Störmann

Abstract	This report demonstrates evaluation results of the three fraud detection systems developed within project ASPeCT for detecting fraud in mobile telecommunications networks. This document discusses the following topics: • A common framework for experiments • First demonstrator experiments • Evaluation of the first demonstrators • Subsequent implementations and trials

Keywords	ACTS, ASPeCT, fraud detection, evaluation

Contents

Executive Summary

This document discusses testing results of the three different ASPeCT fraud detection systems. After describing the concepts (D06 [5]) and the first implementation (D08 [6]) an evaluation of the tools was a natural subsequent task. The evaluation will show that both the concepts as well as the implementation of all three systems are well suited to detect fraud in Mobile Networks.

Since the completion of our three prototypes a lot of convincing demonstrations were held at several locations, partially involving all three systems. We were using a data-sets of fraudsters adapted from the Vodafone TACS system to GSM as well as a two month download of Toll tickets related to Vodafone new subscribers. These different kinds of data-sets, one fraudulent and one almost fraud-free set, are a precondition for assessing the systems' performance.
For the evaluation we choose a systematic testing method (ROC), which is applied for testing in many different scientific areas. This method takes into account that for any classification problem a useful performance statement requires a certain pair of values, in our case the rate of correctly classified fraudsters and the rate of incorrectly classified valid users. The ROC method was commonly used for all three fraud detection tools. Thus it allows comparing the tools in different situations and gives first hints for a integration of all prototypes within a single tool. Additionally, a quite valuable effect of the evaluation was that it produced a lot of advice how to improve each of the systems.

Several statistical curves show how fraud / non-fraud looks, and curves depicting the alarm-levels of users show for each tool how it handles different cases. Each of the tools proves its ability to distinguish between fraud and non-fraud to a good extent.
The next logical step of detecting fraud in the real world is already in progress within ASPeCT. Current batches of new subscribers from Vodafone as well as from Panafon are processed by the fraud detection systems. The most suspicious users found by the systems are passed to the network operators for a thorough investigation. First results from the neural network using unsupervised learning are part of this document.

The specific strengths of the different approaches determined by the tests are quite close to what we expected. The Rule-Based tool which maintains CUPs based on a fixed time interval is good in detecting apparent cases of fraud due to differential and absolute usage at a low rate of false alarms. The unsupervised fraud detection does not rely on prior knowledge of the network's data and is thus very sensitive for new kinds of fraud. Finally, the supervised fraud detection tool is able to automatically learn from test data to achieve a certain abstraction of behaviour patterns of fraudsters and valid subscribers, which allows a good classification during runtime.

We can conclude that the performance results are very promising, especially in view of the few tunings which have been done so far. Additionally, the good results motivate an integration of all demonstrators to an unique system (BRUTUS) in conjunction with a further performance improvement.

Document Control

Document History

This is the second issue of D13 with minor editorial changes made since the first issue.

Changes Forecast

(a) Final version of D13 submitted to commission.	(4-July-97)
(b) Second issue of D13 submitted to commission.	(28-October-97)

Change Control

In conformance with the project Quality Plan [1].

Document Cross References

[1] ASPeCT Quality Plan .
[2] Project Technical Annex (Year 2 Form M4DD).
[3] ACTS AC095, project ASPeCT, “Initial report on security requirements”, AC095/ATEA/W21/DS/P/02/B.
[4] ACTS AC095, project ASPeCT, “Project Trial Plan”, AC095/PFN/W12/DS/P/017/C.
[5] ACTS AC095, project ASPeCT, “Definition of Fraud Detection Concepts”, AC095/KUL/W22/DS/P/06
[6] ACTS AC095, project ASPeCT, “Fraud Management tools: First Prototype.”, AC095/DOC/RHUL/072/WP22/A

Abbreviations and Glossary of Terms

ASPeCT	Advanced Security for Personal Communications Technologies
AU	Analysing Unit
AUA	Absolute Usage Analyser
CU	Controlling Unit
CUP	Current User Profile
DUA	Differential Usage Analyser
EIR	Equipment Identification Register
GUI	Graphical User Interface
GSM	Global System for Mobile Communications
IMEI	International Mobile Equipment Identity
IMSI	International Mobile Subscriber Identity
MA	Master Analyser
MSC	Mobile Services Switching Centre
MSISDN	Mobile Station Integrated Services Digital Network
PABX	Private Automatic Branch Exchange
PDAL	Protocol Data Analysis Language
PDAT	Protocol Data Analysis Tool
PSTN	Public Switching Telephone Network
ROC	Receiver Operating Characteristic
RCF	Roaming Call Forward
SIM	Subscriber Identity Module
SP	Service Provider
TACS	Total Access Communications System
TMN	Telecommunications Management Network
TT	Toll Ticket
UMTS	Universal Mobile Telecommunications System
UPR	User Profile Record
UPH	User Profile History

Introduction

This document reports about the evaluation of the three fraud detection tools, shows results and introduces the subsequent implementation plans of work package 2.2 of the ASPeCT project AC095. A short overview of the following chapters is given here:

- **A common framework for experiments**

This chapter describes the data used for evaluation and introduces the ROC approach of testing and depicting test results, which is commonly used by all fraud detection tools.

- **First demonstrator experiments**

This chapter describes the specifics of testing, experiments with certain subscribers and the overall results depicted in ROC curves for each of the three fraud detection systems.

- **Evaluation of the first demonstrators**

In this chapter experimental results are assessed and the tools are compared to each other.

- **Subsequent implementations and trials**

This chapter introduces the subsequent implementation tasks to be undertaken within wp22. These tasks are partially issued by the evaluation results discussed before in order to improve the tools' performance.

A common framework for experiments

Introduction

Since we have developed the first prototypes within a common framework for the demonstrations, this offers us the possibility of evaluating them within a single framework. This will make comparisons between them easier and will also make an ideal first step towards the integration of the prototypes within a single tool. We briefly describe in this section which experiments were performed and how the performance of the tools were assessed.

Data used for the experiments

To characterize the normal activity on the network, Vodafone provided us with the toll tickets of all new subscribers to their GSM network over the period from July 1, 1996, to August 24, 1996. The prototypes use only six fields from the toll tickets: International Mobile Subscriber Identity, call starting date, call starting time, call duration, B-number, and B-type. The IMSI and the B-number were sanitized. At this point no GSM fraud data were available, but the toll ticket histories to 308 cases of fraud on the Vodafone TACS network were available. This fraud had taken place between July 1, 1994, and January 31, 1995. The same six fields of information are extracted from the TACS toll tickets as from the GSM toll tickets. The IMSI and the B-number are sanitized. We assess the correct operation of the tools using Receiver-Operating-Characteristic curves, which explicitly show the trade-offs between detection of fraudulent users, and production of false alarms. We also illustrate the operation of the tool by analyzing its behavior for a few sample users, namely by looking at the alarm and/or profile histories of two fraudsters (FRAUDSTER 9 and FRAUDSTER 60), and of a regular new subscriber.

Receiver-Operating-Characteristic Curves

To develop statistical methods, researchers often use abstract indexes of performance for which simple optimization methods exist. This allows them to find the “optimal” solution to their problem. However, optimality only refers to “optimality with respect to a given performance criterion”. This means that, in practice, very different performance criteria will be used depending on the application at hand, leading to results significantly different from the results produced by standard statistical methods.

In our case, we focus on finding a performance index that reflects the daily practice of fraud management. The striking feature of fraud detection is the importance of the trade-off between detection of fraudulent users and the production of false alarms. Indeed, we could develop a very conservative system that would generate alarms at the lowest levels of suspicion. But network operators and service providers are, from a commercial point of view, extremely cautious about unduly bothering good subscribers. Moreover, even levels of false alarms that would be considered excellent from a statistical point of view (let us say, one percent of misclassification), would be completely unacceptable in our case because of the large number of users (one percent false alarms for one million subscribers means ten thousand false alarms). Conversely, we could guarantee that we will not generate any false alarms simply by not implementing any fraud detection system. Yet, the burden of loss of revenues caused by fraud makes this solution unattractive. Therefore, the problem of the fraud detection tool will be to find the right balance

between false alarms and correct detection. This optimal trade-off might be different for different operators, different services, or different periods. We thus developed the fraud detection tools to produce a single measure of suspicious behavior each time it receives a new toll ticket. The decision itself comes from choosing an appropriate threshold and deciding to classify a user as suspicious if its activity rises above that threshold at any time of its profile history. A low threshold will guarantee high detection, but will generate many false alarms; while high threshold will guarantee few false alarms, but will detect few fraudulent users.
The Receiver-Operating-Characteristic plots the percentage of correct detection of fraudulent users versus the percentage of false alarms for new users (as illustrated in the coming sections). The index of performance that we need to maximize is the surface under the curve. This is a very practical index of performance, more appropriate to our investigations than standard statistical measures. Such a trade-off curve will give the user of the fraud detection tools control over the fraud detection rate and false alarm rate.

First demonstrator experiments

Tests on a neural net based approach to fraud detection using supervised learning

Introduction

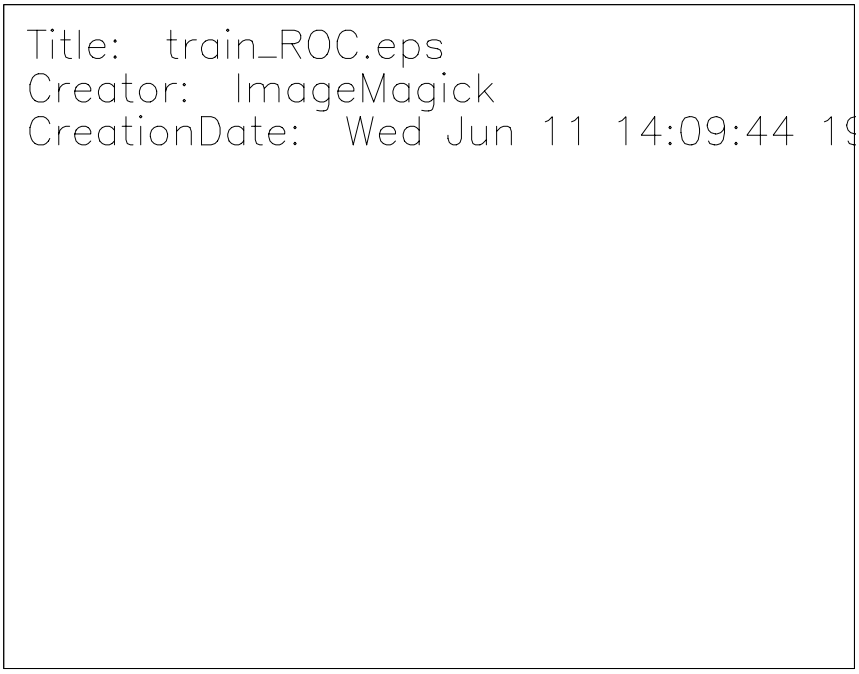
We describe the experiments performed with the fraud detection tool using supervised neural networks. The available data set consisted of the toll tickets related to 308 cases of fraud, and a two-month toll history for 75.000 new Vodafone users. The toll ticket histories relating to fraud were TACS toll tickets translated into GSM format. After describing our training and testing methodology, we assess the performance of our tool at the hand of Receiver-Operating-Characteristic curves, which explicitly show the trade-offs between detection of fraudulent users, and production of false alarms. We also illustrate the operation of the tool by analyzing its behavior for a few sample users.

Training and testing methodology

Of the 308 cases of fraud, only 208 were considered to contain enough information for the development of our tool. The others usually did not have long enough histories so that the tool could not reliably build up the profiles for the differential analysis. We also selected only 300 new users, as we considered it was sufficiently representative of the behavior of normal users. The data was further split up in two. The first half of the fraudsters and of the new users were used for the estimation of the parameters of the neural network model. The other half was left untouched and used only for testing purposes. This splitting of the data set guarantees that the performance evaluated on the test set will generalize well to further cases. We trained by first processing the histories of the 508 training cases to produce 508 profile histories. The architecture chosen for the neural network was a 5 hidden neuron Multi-Layer Perceptron. The parameters of the neural network were adapted using a global optimization procedure to optimize the Receiver-Operating-Characteristic performance criterion described in the next section. Once we had estimated these parameters, we processed the other half of the data (the test data) and we then assessed the performance of the system.

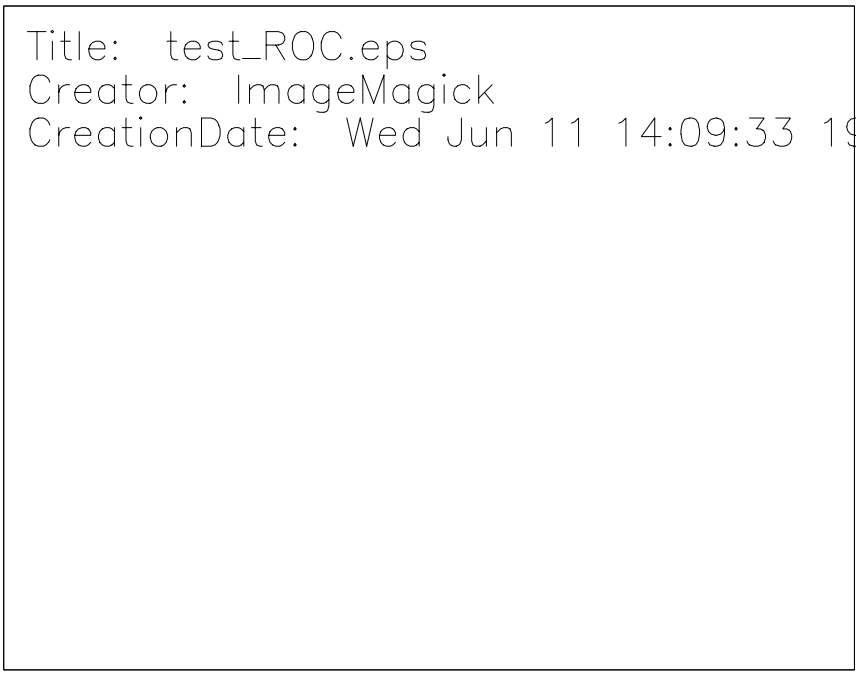
Receiver-Operating-Characteristic Curves

This method allows us to study the trade-off between detection of fraudulent users and the production of false alarms. This trade-off is of great importance. The neural network produces a single measure of suspicious behavior each time it receives a new toll ticket. The decision itself comes from choosing an appropriate threshold and deciding to classify a user as suspicious if its activity rises above that threshold at any time of its profile history. Again, low threshold will guarantee high detection, but will generate many false alarms; while high threshold will guarantee few false alarms, but will detect few fraudulent users. The Receiver-Operating-Characteristic plots the percentage of correct detection of fraudulent users versus the percentage of false alarms for new users. This is depicted in the following figure for the training data. The index of performance that was maximized during training is the surface under the curve.



Percentage of detection of fraudulent users
vs.
percentage of false alarms for new users,
for the training data

We see the same ROC curve for the test data in the following figure. The similarity between the training and the test performance indicate that the model learned the true characteristics of the problem, and not only the characteristics of the few examples available for training. This means that the test performance can be expected to remain similar when the system is tested on further data.



Percentage of detection of fraudulent users
vs.

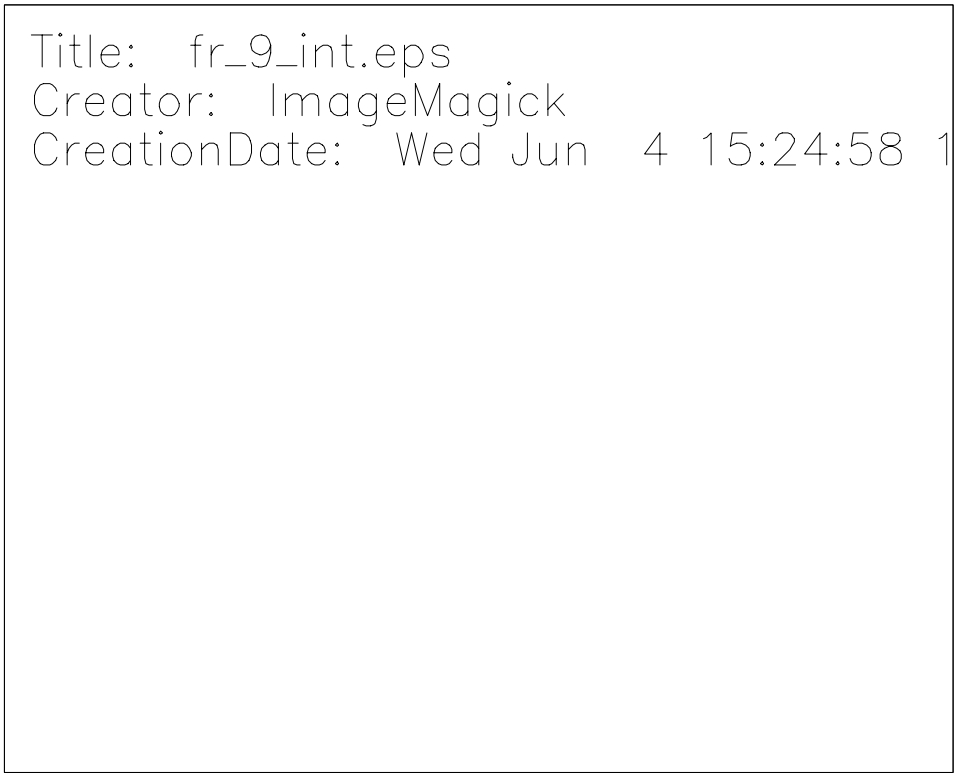
percentage of false alarms for new users,
for the test data

The trade-off between detection and false alarms indicates that we can expect to detect 90% of the fraudulent users for a false alarm rate of 3%. This performance is theoretically very attractive; unfortunately, 3% of false alarms on a customer basis of 1.000.000 customers would generate tens of thousands of false alarms, which is totally unacceptable. A more practical trade-off would bring us at about 60% of detection for 0.3% of false alarms. With such a low level of false alarms, we need to realize that trying to classify fraudulent users as suspicious and new users as non-suspicious might be too dichotomic a choice. In fact, some of the new users might be fraudulent ones too, although in a very small proportion. Although, this proportion is unknown to us, it is expected to lie not much lower than 0.1%. Which means that some of the alarms produced by our system might be worth further investigation. This is currently being handled by Vodafone, as they have already analyzed some of the results of the unsupervised neural network (Section 8.2.4).

Analysis of toll ticket / profile histories

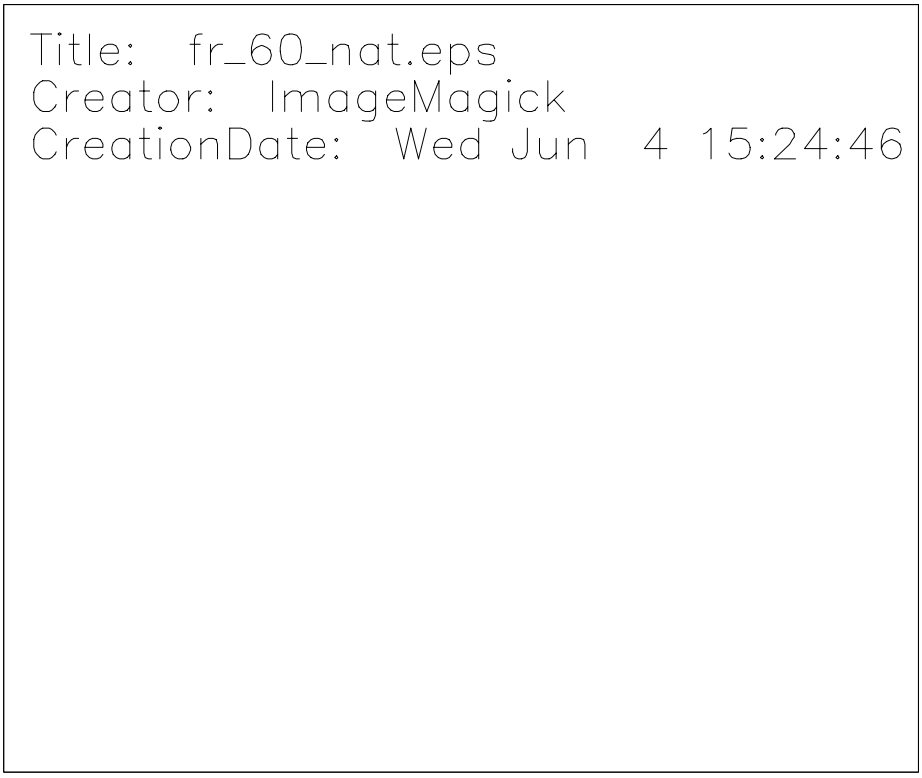
The following plots present the output of the neural networks for two cases of fraud (FRAUDSTER 9 and FRAUDSTER 60) and a normal user. There are eight plots for each users. The first four represent respectively (from top to bottom and left to right) the number of national calls per day, the average duration of national calls, the standard deviation of the duration of national calls, and the duration of each individual call. The last four plots represent respectively (from top to bottom and left to right), the number of international calls per day, the average duration of international calls, the standard deviation of the duration of international calls, and the evolution of the alarm level. The plots for the measurement of national and international quantities contain three curves: the quantity itself, its short-term average and its long-term average.

number of national calls per day, average duration of national calls, standard deviation of the duration of national calls, and duration of individual call

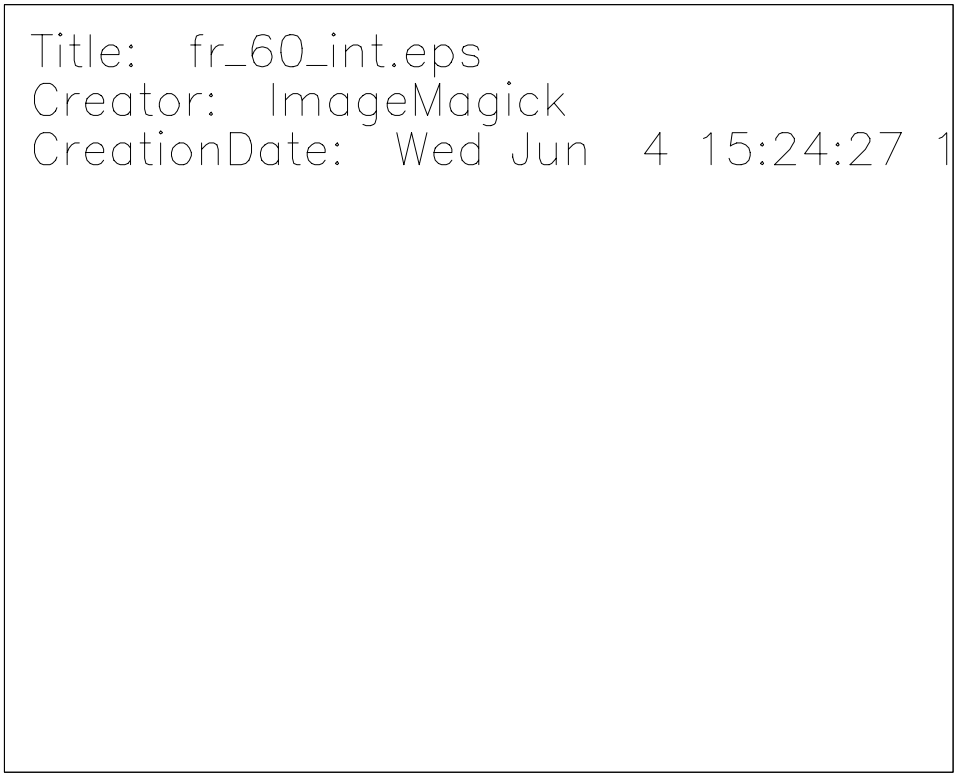


number of international calls per day, average duration of international calls, standard deviation of the duration of international calls, and alarm level

We can notice that the tool reacts as excepted when analyzing FRAUDSTER 9 and sees its alarm level rise (second plot) when the activity of the user suddenly rises (first plot). Similarly, an alarm is raised for FRAUDSTER 60 when its activity suddenly surges (see next two plots.

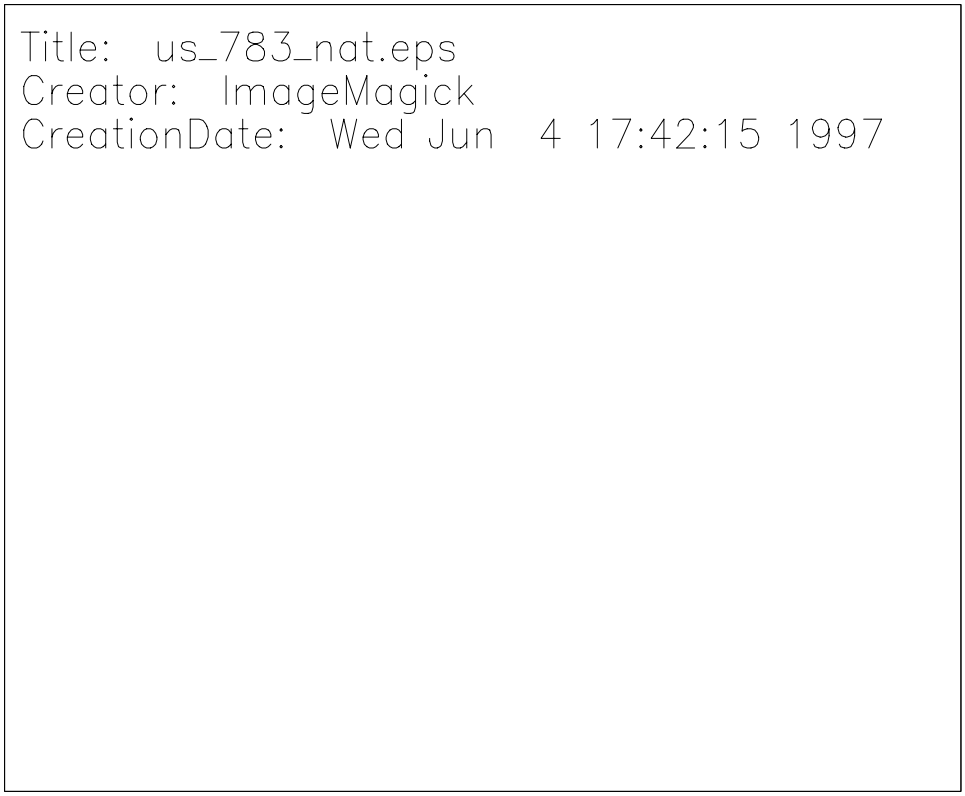


number of national calls per day, average duration of national calls, standard deviation of the duration of national calls, and duration of individual call

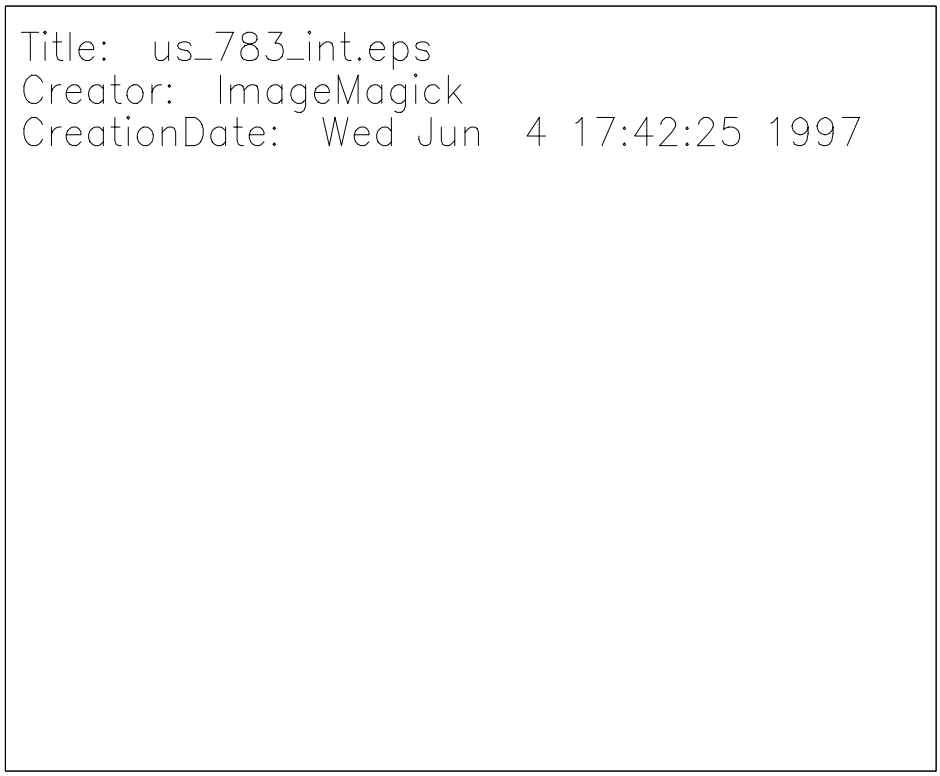


number of international calls per day, average duration of international calls, standard deviation of the duration of international calls, and alarm level

If we now look at a normal user, even if this user has an international activity, the tool still behaves as expected. The alarm level always remain under 0.1.



number of national calls per day, average duration of national calls, standard deviation of the duration of national calls, and duration of individual call



number of international calls per day, average duration of international calls, standard deviation of the duration of international calls, and alarm level

Conclusions

The neural network based on supervised learning thus behaves properly. It is able to track sudden changes or excessive activity by a user and produce alarm levels according to the potential risk of fraud. The neural network was developed using a methodology based on finding optimal trade-offs between detection of fraud and false alarms. This gives good control on the performance of the system. Furthermore, the splitting of the data into a training set and a test set, and the similarity of the performance on these two sets guarantees that the performance results presented here will generalize properly when using the system further.

Tests on a Rule-Based approach to fraud detection

Introduction

In this chapter we discuss evaluation experiments on the Rule-Based based fraud detection tool. The following section introduces alarm histories of a valid subscriber and of a typical fraudster giving an impression what fraud looks like and how the tool behaves. The final section discusses the overall performance of the Rule-Based fraud detection tool.
In order to gain a realistic impression of the tools' performance two different kinds of user-data were processed with the tool. The first set is a subset of Toll Tickets of subscribers to the TACS

network, who have had their service terminated due to a velocity check, indicative of cloning. The second set is a two month download produced by a subset of 2000 new Vodafone subscribers, which is supposed to be relatively free of fraud. While the batch of fraudsters is aimed for measuring the tool's fraud detection rate, the batch of new subscribers determines the tool's false alarms rate. A wide range of these values was determined and is depicted in an Receiver-Operating-Characteristic curve (ROC-curve).

Analysis of Alarm Histories

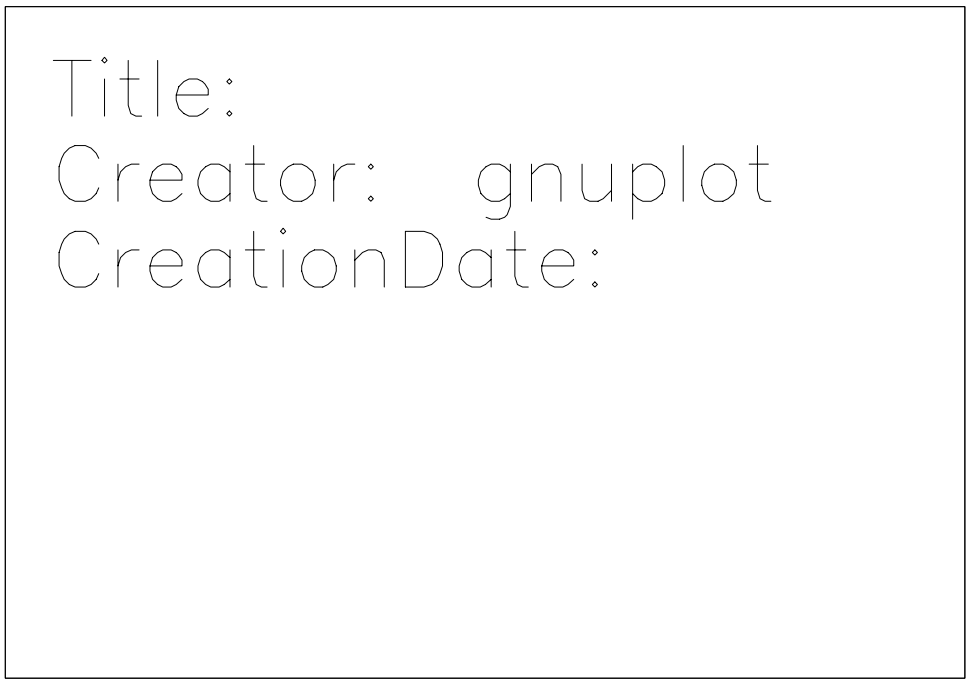
The first step of evaluation was to derive a common alarm level out of the set of rules. The first demonstrator contains roughly three categories of rules, which are rules implementing

- absolute analysis
- differential analysis
- a combination absolute/differential with lower thresholds

For evaluation purposes a common alarm level A across all rules is introduced. Consider i rules of the form: "if value V_i > threshold T_i, then raise an alarm".
Then, the common alarm level A can be defined as: A = max (V_i / T_i), for all rules i. Now, the alarms will raise by one meta-rule "if value A > threshold T_{com}, then raise an alarm". By varying T_{com} we achieve the range of performance values to be shown in the ROC curve.

The applied rules only focus on the user behaviour determined by the most important Toll Ticket features as identified in prior documents (duration and number of national/international calls mainly). As in the other fraud detection tools no overlapping call checks or velocity checks are applied. The rules had been slightly adjusted in relative strength among each other with the full batch of 75000 new subscribers. We did no adjustment based on the fraudulent data to avoid overfitting.

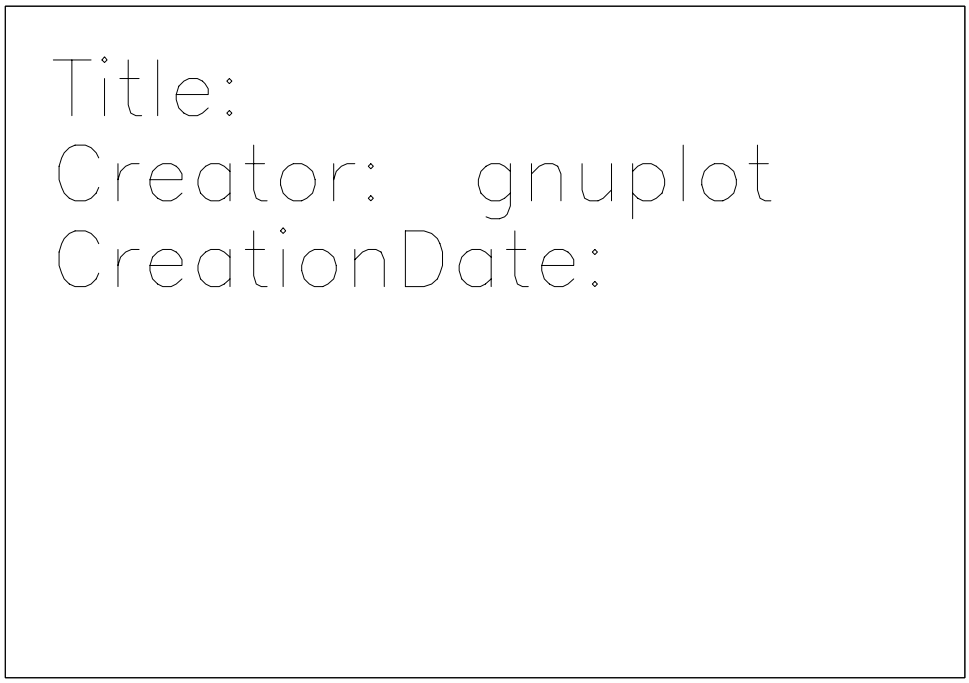
The following two figures show the alarm levels of a new subscriber and a fraudster analysed with the default common threshold T_{com} = 1. In this mode the Rule-Based fraud detection tool found 2 users out of 2000 to be suspicious (alarm rate = 0.1%). Although we assume the batch of new subscribers as fraud-free, there may be doubts whether each of the very few alarms is definitely a false alarm.



Alarm history of new subscriber

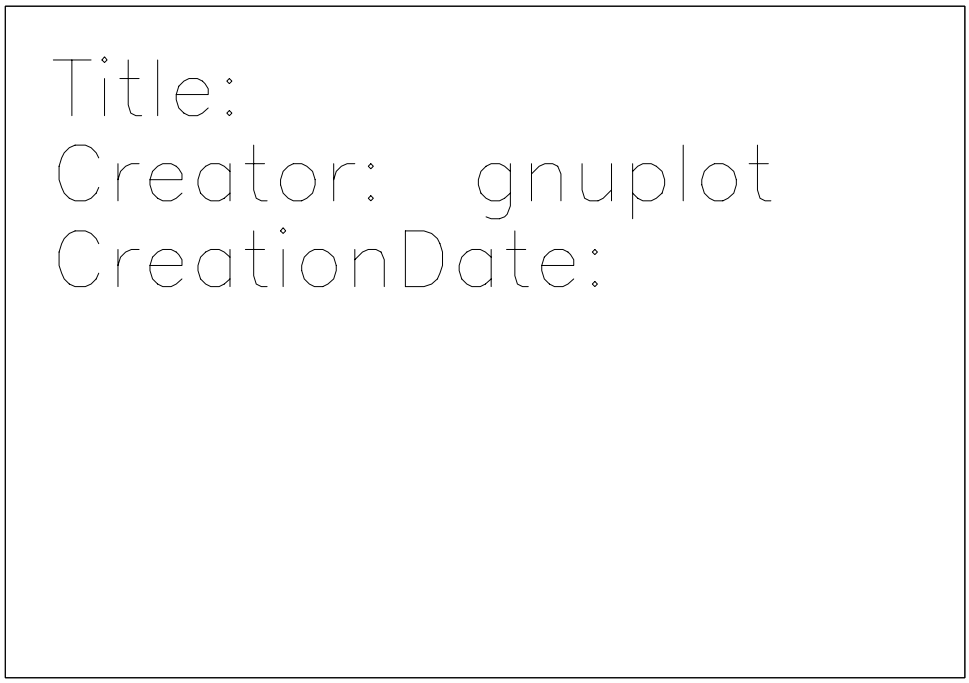
The new subscriber causing the highest alarm level however, seems to be a valid user looking at his alarm history: This user starts some extraordinary activity around day 12 and gets slightly above the alarm threshold for one day. Afterwards however, the alarm level gets back to very harmless values due to a quite normal user activity.

Quite different is the behaviour of the TACS fraudster 009 shown in the next figure.



Alarm history of fraudster 009

Here, the alarm level for several rules is exceeding the threshold for a longer period of time and to a large extent (3 times the threshold). After a short break the fraudster starts a high activity again until he is finally switched off. The three different curves result from different rules: a combined and a differential rule observing the national duration (NatDur.Comb, NatDur.Rel) as well as an absolute rule observing the number of national calls (NrNat.Abs). It is quite interesting to see that each of these curves represents the highest peak for a certain time. At first, the differential rule causes the highest peak, since the gap between current activity and long-term history is quite high at this point. Afterwards the high usage for several days causes the UPH to adapt to this behaviour and the curve falls below the threshold again. However, the combined rule, which also looks for absolute usage stays beyond the alarm threshold. On day 14 an extraordinary high number of national calls leads to the peak of the absolute rule curve. At the end the combined rule has the highest level again. Above all, this example shows that the different rules ideally complement one another.The overall performance of the tool can best be shown as an ROC-curve as already explained in chapter 7.



ROC curve of the Rule-Based fraud detection tool

The highest detection rate determined was 98.6% detected fraudsters, however at the price of 23.9% false alarms, which is impractical. This corresponds to the rightmost point of all three curves and an alarm threshold T_{com} of 0.2. Much more valuable for practical use is a setting of T_{com} around 1. With the default setting T_{com} = 1 we achieve a rate of 83.5% detected fraudsters at 0.1% false alarms. At T_{com} = 1.2 we hit the y-axis at a rate of 78.4% detected fraudsters.
The three different curves result from testing the fraud detection tool with different factors applied at user profile updates. The Rule-Based approach maintains current user profiles on fixed time intervals, i.e. Toll Ticket information is collected for one day in the CUP before updating the UPH at the end of a day. The factor in the figure is multiplied with the old UPH-values before fresh CUP-data is added. A surprising outcome of this test was, that with a quite high factor of 0.96 (i.e. low adaptation) we gain the largest area beyond the ROC-curve i.e. the best performance. In comparison to the old default value of 0.8 we gain a significant improvement of the tool. The drawback of a high factor / low adaptation could be a misclassifying a new high end user as fraudster, when the gap between his activity and the low UPH-values gets too big. However, this apparently did not happen.

Besides, the ROC curves provide a valuable means to improve a Rule-Based system rule by rule. If we apply this method to assess subsets of rules or single rules we can easily identify rules which are useless or which even spoil the system's performance.

Tests on a neural net based approach to fraud detection using unsupervised learning.

Introduction

This section discusses the experiments performed on the neural network fraud detection tool that uses unsupervised learning. These experiments were performed in common with the rule based and supervised system using the limited fraudulent dataset of TACS Toll Tickets, taken from the Vodafone network, which were identified solely through the detection of overlapping calls. We also include feedback from the network operators concerning live IMSI's, identified as exhibiting suspicious behaviour by this fraud detection tool. The goal of this section is to identify the strengths and weaknesses of the unsupervised learning technique and clarify areas in which further work needs to be done, refining theories and enhancing the current implementation. Planned avenues of exploration will be discussed.

Under the original hypothesis the strength of the unsupervised learning system, outlined in previous deliverables, would be the ability to detect new fraud scenarios. One of the possible uses for this would be to feed results to the Rule based FDT and the Supervised FDT enabling them to modify existing rules, add new rules or retrain respectively. This process can be considered as an adaptive critic utilising a management module. A potential weakness of the system was over generalisation leading to poorer performance for certain fraud scenarios.

Output from the unsupervised fraud detection tool are alarm values indicating how erratic the subscriber's behaviour is. These values would normally only appear when their level exceeds a preset threshold essentially determining the number of alarms that we wish to see. Subsequently these alarms would be prioritised for investigation by sorting them into ascending order. For the purpose of the common part of these experiments, this feature has been removed. Instead, alarm levels for each subscriber are produced as each Toll Ticket is processed. The result is a time history of alarm levels for each subscriber that are stored for subsequent statistical analysis.

We conclude that the first prototype of the unsupervised system is capable of performing a differential analysis. Furthermore we have identified through the use of the tool that there are a number of special IMSI's, such as network testing numbers, that need to be filtered out of the datasets prior to processing.

Experiments using the overlapping call data taken from the TACS network.

The unsupervised fraud detection tool works solely on the basis of performing a differential analysis. Statistical user profiles are created by classifying incoming Toll Tickets to prototypes uniformly distributed over a sample dataset taken from a live network. Current User Profiles (CUP's) and User Profile Histories (UPH's) are maintained by the application of two different decay factors to separate probability distributions that represent the short and long term past behaviour of the subscriber. These parameters a and b will be varied relative to each other to test the effect on alarm histories. As will be shown, there appears to be an optimum level for the

difference between a and b such that suspicious activity stands out from normal activity and thus the application of a personalised threshold will be sufficient to trigger the alarm. For the first demonstrator, little performance tuning has been performed. This is part of the current and ongoing activity.

Four different values for a are considered whilst holding b fixed at 0.98. The respective values of a used were 0.5, 0.75, 0.9 and 0.95. The first figure below shows an alarm history for a known fraudster for each of the values a . Notice that between $a = 0.5$ and $a = 0.75$ the significant increase in the alarm value when the fraud begins to occur relative to the period earlier, when only normal activity is taking place.

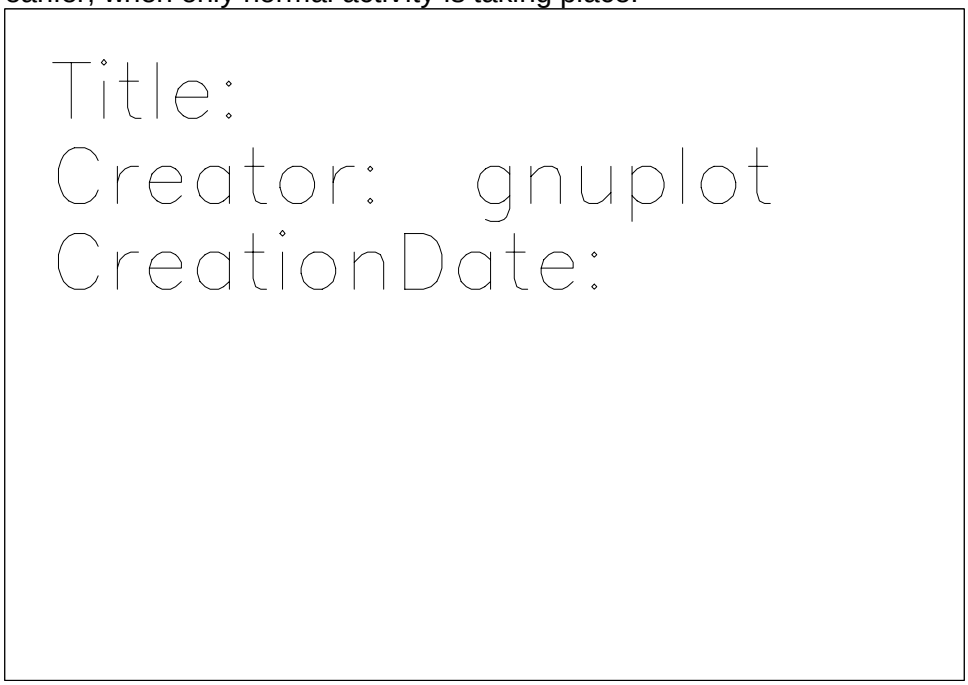


Figure 8.2.1 - Alarm history of a fraudulent TACS subscriber. Fraudster 060

This alarm history shows a clear distinction in the subscribers behaviour before and whilst the fraud is being committed. Other cases in the TACS dataset are not so distinctive. Figure 8.2.2 below show the alarm history for such a fraudster.

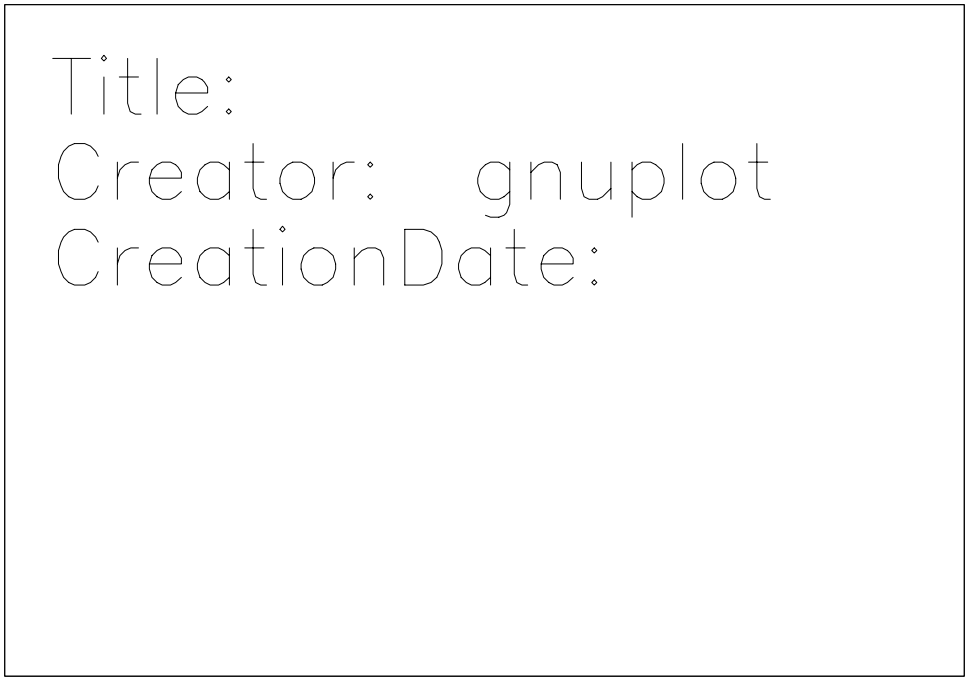


Figure 8.2.2 - Alarm history of a fraudulent TACS subscriber. Fraudster 009

If we compare these alarm histories with those of a new subscriber we notice a number of things.

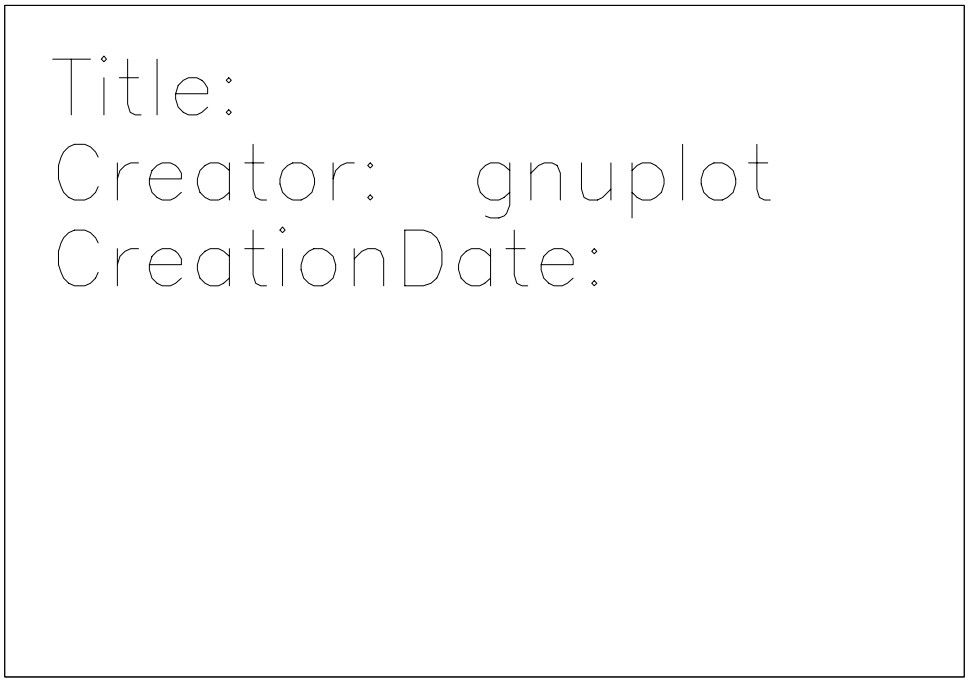


Figure 8.2.3 - Alarm history of new subscriber. New subscriber 1.

Firstly the mean level of the alarm for new subscriber 1 is higher than that of Fraudster 009. Thus new subscriber 1 is displaying more erratic behaviour than Fraudster 009. This is clearly feasible if the cloned phone is being used for personal usage. This presents a problem for a differential analysis that does not yet take into consideration B-Numbers.

Fraudster 060 is clearly distinct from new subscriber 1, also fraudster 009, and can detected quite easily using a personalised threshold. The fact that we cannot distinguish easily between Fraudster 009 and subscriber 1 at present leads to a higher percentage of valid new subscribers raising alarms. This margin can certainly be reduced through introducing B-numbers into the analysis.

One further example of the problems currently faced are when we encounter new subscribers with behaviour that resembles fraudster 060. Figure 4 below shows such a new subscriber where the alarm value steadily increases over a period of time.

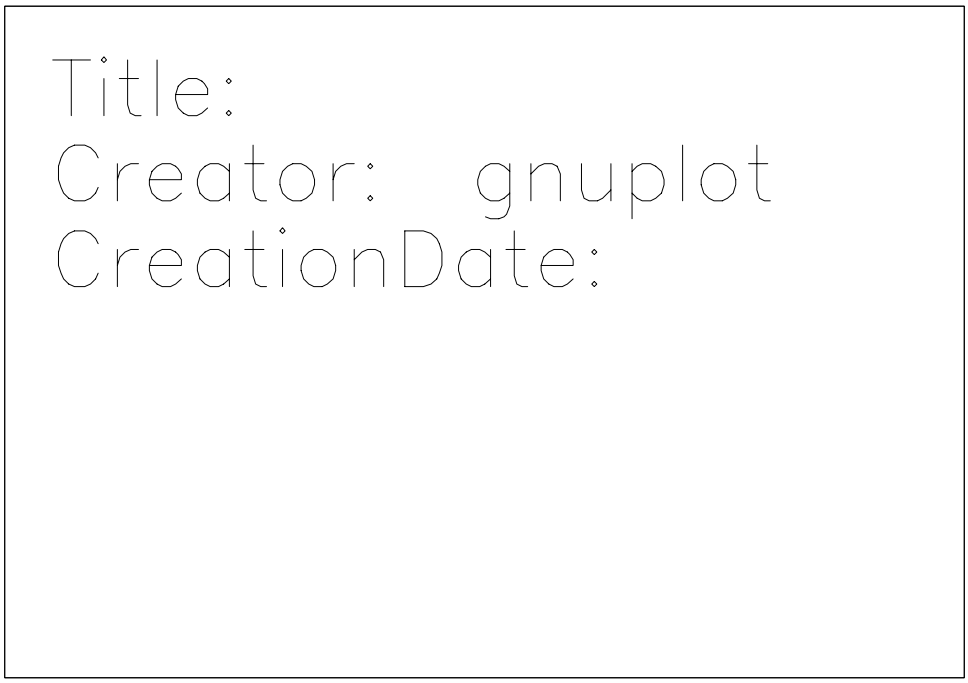
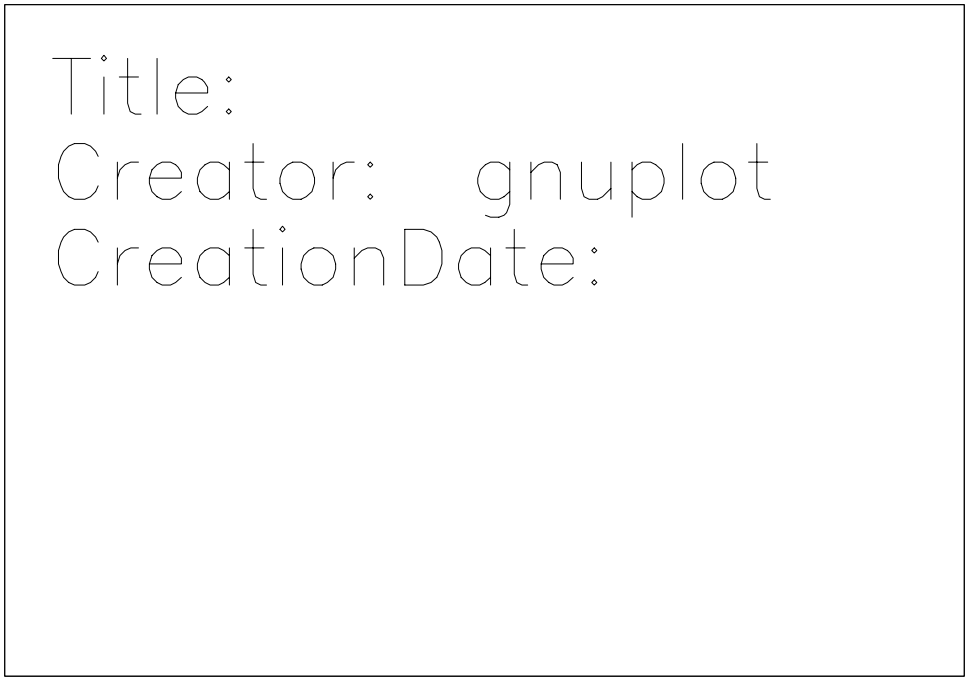


Figure 8.2.4 - Alarm history of new subscriber. New subscriber 2.

Presentation of results.

The following graph shows the performance of the unsupervised FDT applied to the overlapping call data taken from the TACS network. Each graph plots the percentage of the fraudulent IMSI's correctly identified by the fraud detection tool verses the percentage of new subscribers raising alarms at the same threshold.

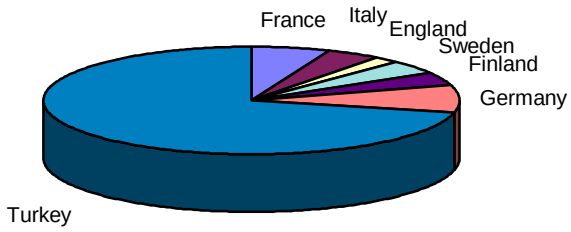

Title:
Creator: gnuplot
CreationDate:

Results of investigation into live IMSI's raising alarms in the unsupervised FDT.

In this section we discuss the feedback from the network operators concerning two lists of IMSI's, detected by the unsupervised fraud detection tool raising strong alarms. The results prove very interesting and demonstrate that the tool is correctly identifying changes in behaviour.

Of the list of suspicious IMSI's submitted to Panafon, consisting of 58 users indicated by the unsupervised fraud detection tool, 9 were Panafon subscribers and 49 were roamers. The composition of the roamers was 3 French, 2 Italians, 1 English, 2 Swedish, 2 Finnish, 4 Germans, and 35 Turks. Note that the IMSIs define the subscriber's home network operator and not his actual nationality. Apparently, subscribers of one Turkish operator[1] had highly suspicous behaviour. None of the subscribers had been disconnected.

Figure 8.2.5 Roamers on the PFN network that raised alarms



A number of the alarms had been raised because of unreleased connections after finishing the calls. These problems are fixed within Panafon before billing the customer. At present, the fraud detection tools do not eliminate such calls from the analysis.

The following table shows the comments from Vodafone concerning the IMSI's that raised alarms when the unsupervised fraud detection tool was applied to their new subscriber data. Again no definite fraudsters, not surprising considering the short period under consideration, but erratic usage is certainly being identified.

IMSI Removed	Fraudulent ?	Results of enquiry / comments
1	NO	Appears to be a test number. Many calls are MO SMS, few MT SMS, few voice. Calls spread out through the day. 26 Voice calls, avg. 92 seconds

[1] Due to the vicinity of Greece and Turkey, roaming is frequent near the border, without actually crossing it.

2	NO	All calls were redirected to a Vodafone Helpline.
3	NO(?)	Calls are redirected to a Vodafone Engineering reserved mobile number. No calls were made from the redirected number in the week, and so it is likely to be an internal test.

4	NO(?)	Calls are redirected to a Vodafone Engineering reserved mobile number. No calls were made from the redirected number in the week, and so it is likely to be an internal test.
5	NO(?)	Calls are redirected to Vodafone Engineering reserved mobile numbers. No calls made from the redirected number, so probably is a valid internal test.
6	NO(?)	All TTs are for SMS services, ~20 MT, rest are MO: phone is resolved to be an SM (Short Message) only mobile.
7	NO(?)	All TTs are for SMS services, all for MT. Vast majority of SMS are within business hours.
8	NO(?)	All TTs are for SMS services, ~50 MO, rest are MT. Phone is resolved to be an SM only mobile
9	NO(?)	All TTs are for SMS services, all for MO. All between 8am and 4pm. Phone is resolved to be an SM only mobile.
10	NO	IMSI is produced by operator switches - no specific subscriber but a system property.
11	NO	Only a few international calls.
12	NO	Only ~50 calls made to international destination - all to Lebanon, avg. 60 seconds
13	NO	Unresolved but appears to be non fraudulent.
14	NO(?)	No international calls; all mainly to one number
15	NO	Only a few international calls
16	NO(?)	Foreign subscriber; unusual call start time distribution.
17	NO(?)	Overlap in TT - explained by mobile being busy and incoming call being diverted.
18	NO	No international calls, just national calls with +44 prefix.
19	NO	17 international calls, unlikely to be fraudulent.
20	NO(?)	OVERLAPS IN TT - explained by mobile being busy and incoming call being diverted.
21	NO	Unresolved but appears to be non fraudulent
22	NO(?)	Call start time distribution appears to be consistent with a business number.
23	NO(?)	Appears to be consistent with a business number.
24	NO(?)	Looks like a business front end. Redirected out of hours.
25	NO(?)	All calls redirected - call times seem to be concurrent with business hours.
26	NO(?)	Calls redirected - Vodafone Engineering reserved number. Calls look machine generated.
27	NO(?)	Calls redirected - Vodafone Engineering reserved number, however there are a lot of different calling parties.
28	NO(?)	Calls redirected - Vodafone Engineering number. As previous. 2
29	NO	IMSI is produced by operator switches - no specific subscriber but a system property.
30	NO	Calls redirected to a Vodafone Helpline or similar.
31	NO	All international calls back to Germany.

32	NO(?)	Just an irregular call start time distribution. Only a few international calls
33	NO	Only a few international calls, most of which are back to the home country. Most call destinations are to British numbers.
34	NO	All international calls back to home country.

35	NO(?)	Most suspicious of all, but still within realms of a business user making some overseas calls. Interesting averages - 14mins for international, 4mins for national.
36		** See end for comments on this subscriber **
37	NO	Foreign subscriber, lots of calls back to Greece.
38	NO(?)	17 international calls, unlikely to be fraudulent.
39	NO(?)	OVERLAPS IN TT - explained by mobile being busy and incoming call being diverted.
40	NO(?)	Foreign subscriber, mainly international calls made back to own country.
41	NO(?)	Foreign subscriber; unusual call distribution.

"Interestingly enough, subscriber 36 is no longer on the network (on that number...). A brief history is as so:
- Connection in 1994
- Admin. bar in mid 1996 - subscriber had gone roaming, made a huge bill of £2400 since the last
  invoice (ignorance of charging systems), and so was temporarily barred. The bill was paid, and the bar lifted.
- Early 1997, subscriber decided to terminate their contract with Vodafone, and left the network."

From the results we see that there are a number of important reserved numbers that need to be identified so that they can be filtered from the analysis. Identification of all these numbers will be requested from the network operators.

Conclusion

In this chapter we have discussed experiments performed using the unsupervised FDT. The main distinction between this tool and the rule based and supervised tool is that it has no prior knowledge of the data that it operates on. This makes the tool very flexible and with a great potential for detecting new fraud scenarios. Clearly this is important for GSM where little fraud other than Subscription Fraud currently exists.

Early results from the first demonstrator show that the system is capable of performing a differential analysis. The tool performs surprisingly well on data not particularly well suited to the environment in which the tool would normally operate, i.e. over longer periods of time.

Much work needs to be done refining the current implementation. This will be helped by the arrival of more Toll Tickets in the near future. In the following sections we discuss in more detail the relative merits of each of the systems and also the scope for future work and integration of the systems.

Evaluation of the first demonstrator

In section eight we have described the experiments performed on the prototypes of the three fraud detection techniques namely, Rule-Based, Unsupervised Neural Network and Supervised Neural Network.

The experiments were performed in common using a limited set of fraudulent Toll Tickets, taken from the Vodafone TACS network, and a two months download of new subscriber data taken from the Vodafone GSM network.

The experiments have demonstrated that all three tools have the capability to distinguish fraudulent activity from normal usage to a certain extent. As we originally envisaged, each tool has its own strengths and weaknesses as shown by their differences in performance on the demonstrator data. The results of the experiments conformed with the original hypothesis however and are on the whole very pleasing. Little performance tuning has been undertaken at this stage and yet we are getting results that are comparable with other fraud detection tools currently in use in other industries.

It is clear that the strength of the Rule-Based fraud detection tool lies with its flexibility to add, remove and update rules as desired. Also once a rule has been correctly derived, it tends to be very effective. The performance of the system on the TACS fraud data was exceedingly good. One reason for this is the possible strength of the profiling technique. With the fraudulent TACS dataset, the tendency was to see immediate increases in call frequency and a large change in call durations. With the Rule-Based fraud detection tool, the UPH is updated, based on the CUP, once per day. This means that the on calculating the difference between the profiles, larger overall changes will be seen than with the profiling techniques that updates the UPH after every Toll Ticket.

The strength of the Unsupervised fraud detection tool lies with the fact that it does not require fraudulent data for training. Instead the tool defines the boundaries of acceptable usage on an individual subscriber basis. On the overlapping call data it did not perform as well as the Rule-Based or supervised tool, however this was to be expected. Unlike the other two tools, the unsupervised system has no prior knowledge of the data. On the live network data, the tool managed to detect a number of irregularities, mainly network testing numbers and roamers with odd behaviour. These results are considered of value and will be useful for filtering future IMSI's that raise alarms and are already known about.

The supervised fraud detection tool proved very effective at learning the behaviour patterns of the fraudulent TACS users. The ability of the Neural Network to generalise from training is of utmost importance in this analysis. The experiments have demonstrated that this does indeed occur in this case. Performance figures are approaching that of the rule-based tool and there exists still more potential for increased performance after further tuning.

Potentially the most promising development will be the combination of the three tools into one hybrid system which will be known as BRUTUS (**B**-number and **RU**le-based analysis of **T**oll Tickets utilising **U**nsupervised and **S**upervised Neural Network technologies). The plans for this hybrid system will be outlined in the following section.

Subsequent implementations and trials

Future work will be concentrated in two main areas, namely the improvement of the individual FDT modules and their unification into a hybrid system known as BRUTUS (**B**-number and **RU**le-based analysis of **T**oll Tickets utilizing **U**nsupervised and **S**upervised Neural Network technologies). In addition to the Rule-Based, Supervised and Unsupervised modules, BRUTUS will have a further module that performs a B-number analysis. A B-number analysis should improve the detection process by profiling call destination characteristics for a number of calls made on a regular basis. If the subscriber totally fails to call any of the common numbers in his profile, this information could add weight to any sudden increase in abnormal activity that had been detected, thus increasing confidence in the diagnosis. Furthermore some international destinations are typical for fraudulent calls and are considered to be 'hot' destinations. We aim to provide a flexible method for weighting calls to different countries as part of the B-number analysis tool.

Improvements to the Rule-Based system will concentrate on ways to improve the detection ratio. Firstly this will be done by adding new heuristics, as rules, and improving the profiling technique. We hope to develop a slightly more abstract way of generating new rules automatically via an adaptive critic.

The Unsupervised Learning system will concentrate on improving its prototyping technique so that the prototypes extend into interesting regions of the behaviour space where fraudulent activity is more prevalent.
New decaying techniques will be investigated in addition to current fading procedures so that the UPH is not updated so frequently, a feature which has been a notable result from the Rule-Based tool.

The Supervised learning tool will concentrate on utilizing more features for input to the analysis. Two of these features will be the output from the Rule-Based and Unsupervised Learning tool. The first stage in this procedure must be to enable communications between the different modules to form BRUTUS. We will do this through the tagging procedure described in the following section.

Inter Module Communications.

We intend to implement a scheme for sharing of information between the fraud detection modules. Individual modules should forward information that they receive from any other module and add tagged information of their own should it be required. The data should be fully human-understandable at all times. It should be structured as a sequence of tag/value elements.

Tags are four-printable-symbol strings.
Values are arbitrarily long printable-symbol strings.
The size of a Tag label should be fixed.
Blank spaces are used to separate tags and values.

The input behaviour of each tool should be the following. When it reads a string, it scans it for the tag of the fields it wants to use and extracts the corresponding values, overlooking the tag/value pairs it does not use for its own processing. It is safe to assume that the first six elements (twelve

fields) correspond to the six toll ticket fields used in the first demonstrator. The output behaviour should be the following. The tool copies the string it received at its input directly to its output, and adds tag-value pairs for all the information it wants to output (for example, for use by the monitoring tool). Writing to standard output should only happen at one place in the code, so that the structure of output can be updated easily.

Example:

The output of the Toll Ticket simulator might look as follows.

TMSI F23415124546303b2d224c63 TCSD 19960716 TCST 220038 TCDR 000693 TBNB FFFFFFFFFFFFFF30198672b641014 TBTP 01 TSDN 0016 TSTS 079238 TBZC 03
TMSI F23415124546303b2d224c63 TCSD 19960716 TCST 221856 TCDR 000031 TBNB FFFFFFFFFFFFF017433333d571a4f TBTP 00 TSDN 0016 TSTS 080336 TBZC 00
TMSI F23415124546303b2d224c63 TCSD 19960716 TCST 222015 TCDR 000367 TBNB FFFFFFFFFFFFF017433333d571a4f TBTP 00 TSDN 0016 TSTS 080415 TBZC 00
TMSI F23415140807624d312f4b4c TCSD 19960716 TCST 224913 TCDR 000003 TBNB FFFFFFFFFFFFFFFFFFF4646250c79 TBTP 00 TSDN 0016 TSTS 082153 TBZC 00
TMSI F23415137f381c3f526f710a TCSD 19960717 TCST 074916 TCDR 000001 TBNB FFFFFFFFFFFFFFFFFFF4646250c79 TBTP 00 TSDN 0017 TSTS 028156 TBZC 00
TMSI F2341513363e667947231933 TCSD 19960717 TCST 080242 TCDR 000023 TBNB FFFFFFFFFFFFFF301423d5c494b49 TBTP 01 TSDN 0017 TSTS 028962 TBZC 03

This would be sent to the fraud detection tool. But the tool might not need all these fields. Let's say that instead of using the B-type, it prefers to use the zone code TBZC that gives the region of the world the call was made to. And it does not use the TSDN and TSTS fields. It further processes the information, possibly producing an alarm. At the output, it copies what it received at the input plus all the information it finds relevant (here, the information about the alarms). The output could look as follows:

TMSI F23415124546303b2d224c63 TCSD 19960716 TCST 220038 TCDR 000693 TBNB FFFFFFFFFFFFFF30198672b641014 TBTP 01 TSDN 0016 TSTS 079238 TBZC 03 SALR ALARM SALV 0.87
TMSI F23415124546303b2d224c63 TCSD 19960716 TCST 221856 TCDR 000031 TBNB FFFFFFFFFFFFF017433333d571a4f TBTP 00 TSDN 0016 TSTS 080336 TBZC 00 SALR NOALR SALV 0.37
TMSI F23415124546303b2d224c63 TCSD 19960716 TCST 222015 TCDR 000367 TBNB FFFFFFFFFFFFF017433333d571a4f TBTP 00 TSDN 0016 TSTS 080415 TBZC 00 SALR NOALR SALV 0.33
TMSI F23415140807624d312f4b4c TCSD 19960716 TCST 224913 TCDR 000003 TBNB FFFFFFFFFFFFFFFFFFF4646250c79 TBTP 00 TSDN 0016 TSTS 082153 TBZC 00 SALR NOALR SALV 0.12
TMSI F23415137f381c3f526f710a TCSD 19960717 TCST 074916 TCDR 000001 TBNB FFFFFFFFFFFFFFFFFFF4646250c79 TBTP 00 TSDN 0017 TSTS 028156 TBZC 00 SALR ALARM SALV 0.98
TMSI F2341513363e667947231933 TCSD 19960717 TCST 080242 TCDR 000023 TBNB FFFFFFFFFFFFFF301423d5c494b49 TBTP 01 TSDN 0017 TSTS 028962 TBZC 03 SALR NOALR SALV 0.07

Tags from the toll ticket simulator start with a T, tags from the supervised neural network start with S, tags from the unsupervised neural network start with U, tags from the rule-based system start with R, tags from the B-number analysis start with B.