Telecommunications Fraud: Problem Analysis - an Agent-based KDD Perspective

Eugénio Rosas and Cesar Analide

Department of Informatics
University of Minho
Braga, Portugal
eugeniorosas@gmail.com, analide@di.uminho.pt

Abstract. Telecommunications fraud is a problem that affects operators all around the world. Operators know that fraud cannot be completely eradicated. The solution to deal with this problem is to minimize the damages and cut down losses by detecting fraud situations as early as possible. Computer systems were developed or acquired, and experts were trained to detect these situations. Still, the operators have the need to evolve this process, in order to detect fraud earlier and also get a better understanding of the fraud attacks they suffer. In this paper the fraud problem is analyzed and a new approach to the problem is designed. This new approach, based on the profiling and KDD (Knowledge Discovery in Data) techniques, supported in a MAS (Multiagent System), does not replace the existing fraud detection systems; it uses them and their results to provide operators new fraud detection methods and new knowledge.

1 Introduction

1.1 Motivation

Following the exponential growth in the telecommunications sector in the end of the past century, the telecommunications operators face a new challenge: fraud. It is not only a risk, but a highly organized global business, that affects operators all over the world. In order to realize the severity of this problem, CFCA - Communications Fraud Control Association [1] (the Premier International Association for revenue assurance, loss prevention and fraud control) published some statistics stating that the annual global fraud losses in the telecoms sector are now between US$54 billion and US$60 billion, an increase of 52% since 2003.

Operators know that this is a problem that cannot be extinguished, so the approach to this problem is to minimize the losses. In order to do so, the operators have developed or acquired fraud detection systems. However, these systems were designed specifically to detect a fraud situation, this is, after it has happened. By detecting a fraud situation the operator prevents further losses, but since the fraud has already occurred, the operator has to support the fraud inherent losses. This is an important issue to the operators, because since they cannot totally eradicate fraud, it is very important to detect the fraud as soon as possible in order to minimize the losses.

1.2 Objectives

This paper has the following main objectives:

- Expose the fraud problem within the scope of telecommunications, enumerating the main causes for telecommunication fraud and the impact on the operators;
- Analyze the actual fraud detection solution, explaining the methods used by the fraud solution to detect fraud and analyzing how the solution can evolve;
- Define a new approach in order to evolve the actual fraud detection solution, defining new methods;
- Propose a solution model, which supports the methods previously defined. We will show that a MAS is a suitable approach to build a model to the solution of the problems addressed here.

1.3 Structure

The following chapter has a more detailed analysis to the fraud problem. After this analysis, an approach to solve the problem is suggested, presenting the data structure the solution is based on and the techniques used. The paper ends with a conclusion and an analysis to the work done.

2 Problem Analysis

2.1 Fraud causes

It is not possible to enumerate completely and exhaustively all the existing fraud types. Due to the constant evolution of technology existing fraud types are adapted and new fraud types are developed all the time. However, there is a set of the major fraud causes that are the ones of most concern to the telecommunications operators at the time. These are:

- **Subscription fraud** - one of the most common fraud types along with the SIM cloning. The fraudster obtains the service from the operator with no intention to pay for it, using a false identity. The damage this fraud type causes depends on the intention of the fraudster: using the service for personal use until he is detected; on a more sophisticated level, the fraudster can use the service in order to profit from the use of it.
- **Bypass fraud** - deprives the terminating operator of interconnect termination fees for incoming international calls. This is usually done using VoIP technology to bypass international calls.
- **SIM cloning** - a fraudster clones an existing normal SIM card. The software to clone SIM cards is available on the internet, so if a fraudster has physical access to a SIM card all he needs to clone it is a PC and a card reader. This is considered the most common fraud cause of all [10].
- **Internal fraud** - implies action of internal staff of the operators. Typically, operator employees with knowledge and access to the information systems, handle information in order to benefit a third party, for instance: giving free minutes, changing account settings.

2.2 Fraud detection

Over the years, telecommunications operators have developed or acquired technology in order to indentify fraud situations. This technology is based in a set of methods specifically designed to detect fraud. Some of the most common methods are:

- **High Usage** - measure the amount of traffic generated by each SIM card; detect SIM cards that generate high amounts of traffic in the operators' network.
- **Calls collision** - monitor the traffic in a time dimension for each SIM card; detect overlapped events generated by the same SIM card.
- **IMEI/IMSI stuffing** - mapping the SIM cards (IMSI) to the devices, cell phones for instance, they are used in (IMEI); detect devices that use several SIM cards.
- **Call velocity** - monitoring the traffic in a time and geographic dimensions for each SIM card; unlike the calls collision method, this method's gold is not to detect overlapped events, but physically impossible. For instance, the same SIM card has a call that ends at 2:00am in Braga and another one that starts at 2:10am in Faro (more than 600Km away). Despite the events are not overlapped, it is physically impossible to travel the distance in 10 minutes.
- **Ratio** - monitor the services used by each SIM card; detect SIM cards that use services (for instance: Voice, SMS, MMS, data) disproportionally.

These methods are implemented and thresholds are set by the operators' analysts. The methods are continually monitoring the network traffic and keep statistics for a given time period. If the threshold is reached, a possible fraud situation is detected. Then the analysts decide on each situation if this is a real fraud situation and what actions to take.

Still, the operators are struggling against a huge problem: even if a fraud is detected, all the damage has already been done. The methods to detect fraud that currently exist are reaction-based; they can only detect a fraud after it has already happen, avoiding further damage, but still supporting all the inherent damage to the situation.

Another major issue in the fraud detection systems is the lack of a knowledge base. Because of the complexity of the systems it is very complex to find a structure that stores all the fraud situations previously detected. This knowledge base could be very useful in order to the operators better understand the fraud attacks they suffer and retrieve valuable information.

2.3 Defining an approach

The new approach defined is complementary to actual fraud solution: it does not replace it, but uses the data processed by the fraud solution and the fraudsters detected by it in order to evolve the solution to a new level. The data processed will be used to build a profile for each subscriber, containing individual and behavior features. When a new fraudster is detected by the fraud solution, the profile will be used for: the identity features will be used to detect an attempt from the fraudster to re-enter the operator network; the behavior features will be used to detect other subscribers that have a similar behavior and therefore are likely to commit fraud.

In order to solve the problem previously detected that affects the telecommunications operators, the following goals were defined:

1. Define a method that monitors the operator network traffic and builds profiles based on the SIM cards actual usage. These profiles are composed by identity and behavior attributes.
2. Define a method that monitors the operator network traffic and detects fraud suspects based on a behavior comparison using the behavior attributes of the profiles of fraudsters previously detected by the fraud solution.
3. Define a method that monitors the operator network traffic and detects previously blocked fraudsters attempting to re-enter the network, based on an identity comparison using the identity attributes of the profiles of fraudsters previously detected by the fraud solution.
4. Define a knowledge base structure, capable to support information retrieved from all the known fraud cases.

Notice that, as it was mentioned before, this new approach is complementary to actual fraud solution, it does not replace it, but works cooperative with it in order to detect fraud suspects as soon as possible.

3 Solution proposal

The solution proposed to solve this problem is based in two main techniques: profiling and KDD (Knowledge Discovery in Data). These techniques will be implemented in a MAS (Multiagent System). After an initial section where the data structure that will be used as input for the solution is explained, the MAS proposal is presented, followed by a section where each agent of the MAS is detailed.

3.1 Data structure

In a telecommunications company the operator keeps record of every event processed by the system. These events are recorded in CDRs (Call Detail Records), generated automatically and are used for billing purposes. Each CDR has information regarding a set of events, voice calls or SMSs, for example. Typically, the CDR is a text file containing information structured by a pre-defined set of ordered fields separated by a pre-defined character. Each line of the CDR file is an event processed in the operators system. The structure of the CDR (the number of fields, the order of the fields and the separator character) is defined by the telecommunications company, so the CDR structure varies from operator to operator. However, there is a set of fields that, due to their importance, for billing and rating purposes, are usually common to all CDR structures:

- **A Number** - identifies the originator of the event;
- **B Number** - identifies the receiver of the event;
- **Event Date** - the date the event started;
- **Event Type** - identifies the type of the event, for example: 1 (Voice), 2 (SMS), 3 (MMS), 4 (Data);
- **Event Amount** - measure of the event, for example, in a voice call the event amount is 124 seconds, in a SMS the event amount is 45 characters;
- **Cell ID** - identifies the network cell that processed the event.

The information contained in the CDRs will be the input for all the future work. The study of the contents of CDRs is not a novelty. They were first created with billing purpose, but know they are used with different purposes of great importance to the operators, like discovering user communities [14], for instance.

3.2 Multiagent System

Distributed Artificial Intelligence (DAI) was a subfield of Artificial intelligence research dedicated to the development of distributed solutions for complex problems [2]. These days, DAI has been largely supplanted by the field of Multiagent Systems (MAS). The main purpose of the MAS is the study, construction and application of Multiagent Systems, that is, systems in which several interacting, intelligent agents pursue some set of goals and/or perform some set of tasks [15]. Some motivations that led to the development of the MAS: complexity and dimension of the problems; problems geographically and/or functionally distributed; information and knowledge disperse; multiple systems interconnection; parallelism, robustness and scalability.

The MAS implemented to support the solution proposal is a **closed MAS** [16][12], where the architecture design is static, with all the agents and functionalities pre-defined. In this closed MAS the agents communicate using a common language, each agent is developed as an expert in his functionality and they all work and cooperate together in order to achieve a main goal.

The coordination of the MAS is **cooperative** [11]: the agents do not compete; they cooperate in order to achieve a main objective. The organization is **flat** [11]: each agent is an expert in an area, there is no agent commanding other agents and all agents have the same importance. The communication between the agents is **direct** [11]: there is no agent or middleware between two agents communicating with each other, they communicate directly.

The following image illustrates the MAS architecture and the process flow between the MAS and the fraud solution.

As for the architecture, the MAS is composed by three agents:

- the **Profiling Agent** is responsible for integrating the CDR files and building the profiles (with identity and behavior features) for the subscribers;
- the **Detection Agent** is responsible for detecting new fraud suspects;
- the **KDD Agent** is responsible for processing the profiles of known fraudsters and enrich the knowledge database.

As for the process flow between the MAS and the fraud solution, the flow of information is represented by numerated arrows:

- 1 and 2 - input for MAS from the fraud solution;
- 3 and 4 - internal MAS information exchange;
- 5 and 6 - output of the MAS to the fraud solution.

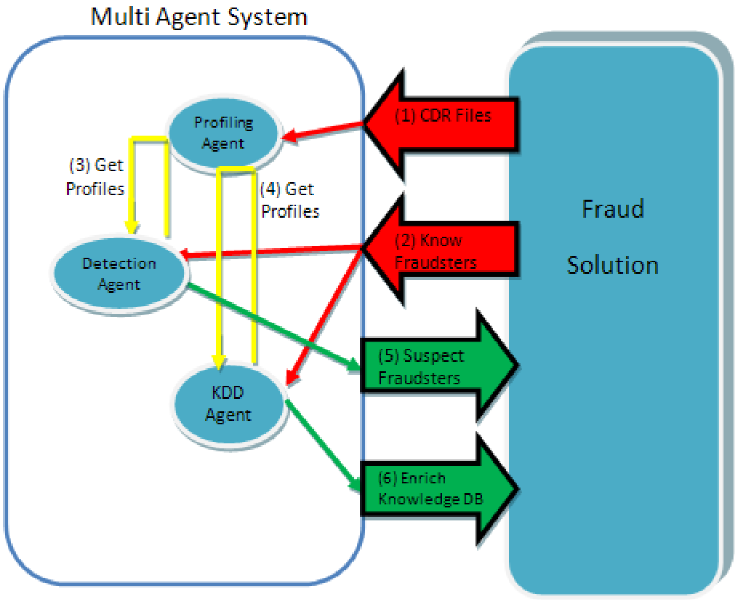

Fig. 1. MAS - Architecture and Process Flow

In the first phase, the fraud solution redirects all the CDR files contents for the MAS (**step 1** in the figure), where the profiling agent uses this content to build the profiles for the subscribers. In this phase the profiling agent should extract from the CDR files contents the necessary information to in order to enrich the identity and behavior features of profiles.

In a second stage, when the fraud solution detects a new fraudster, based on some of the methods previously explained, the fraud solution indicates the MAS that a new fraudster was detected (**step 2** in the figure). Then, this information is used by two agents with different purposes:

- the detection agent will use this information in order to retrieve from the profiling agent the profile of the fraudster and use the profile identity features to detect if the same subscriber tries to re-enter the operator network and the profile behavior features detect other subscribers that have a similar

behavior; the fraud solution is then warned of the suspects that this agent detects (**step 3** in the figure).
- the KDD agent will use this information in order to retrieve from the profiling agent the profile of the fraudster and use the profile behavior features and enrich a knowledge database containing all the detected fraudsters profiles (only the behavior features) in order to try to retrieve significant information (patterns, similar behaviors) from this database. The results of this enrichment is then passed to the fraud solution (**step 4** in the figure), so that the fraud analysts have access to this information.

3.3 Profiling

Profiling is an auxiliary technique for criminal investigation. It fits in the Forensic Psychology domain. Profiling consists in a process of individual features inference, usually individuals responsible for criminal actions [4]. The profiling technique should be used as an extension of the criminal analysis [9], elaborating criminal profiles based on previous work. The mains idea to retain is: profiling complements previous work, it does not replace it.

Nowadays, profiling is a very used technique, implemented in police forces all around the world. Experts in this particular technique, like McCrary [7] or Wrightsman [17], point out the fact that the results of the use this technique make excellent prediction factors.

Even though the profiling technique was developed under the Forensic Psychology domain, it is a technique that is being used in the computer information systems, with various purposes. There are applications of profiling in order to detect social networks [6] [5], large scale behavior analysis [13], security applications [8] [18] and data transaction prediction models [3].

3.4 Applying profiling

Applying the profiling technique using the information contained in the CDRs, it is possible to build a profile of the operator users. The profile is divided in:

- **Behavior profile -** profiling about the user behavior in the network: which services the user uses and in which ratio, the periods of the day, week month he uses the services.
- **Identity Profile -** profiling about the user identity: the user's social network and the community he belongs to, geographical location.

For each network user, the profile should be continuously updated, and the evolution of the profile should also be recorded, because the evolution is by itself a profile input. As it was previously explained, the profiling technique should be used as an extension of the criminal analysis. The same applies in this case. The profiling does not substitutes the fraud detection methods previously explained; it uses and depends on them. The fraud detection methods are vital to detect the fraud cases and identify fraudster users. Only then, the profiling technique results will be used. Once detected a fraud situation and identified a fraudster, the fraudster profile will be used to:

- Enrich a knowledge database. Afterwards, it will be necessary to define methods that allow analysts retrieve important and relevant information from these databases. This information will give the operators the indicators regarding the fraud attacks they suffer. Who, when, where and how fraudsters attack the operators are questions that will have some answers thanks to this knowledge database.
- The identity profile of the fraudulent user should be saved and used to identify future attempts of the same user to use the network services.
- The behavior profile, including the evolution history, will be compared to other users, in order to detect users that are likely to commit fraud in a near future.

3.5 Agent-Based Knowledge Discovery in Data

In the past twenty years, agents and KDD (Knowledge Discovery in Data) have emerged separately as two prominent, dynamic and exciting research areas [19]. In recent years, an increasingly remarkable trend in both areas is that integrating agents into distributed data mining systems in order to extract knowledge from data in a faster, efficient way.

The KDD process consists in:

- **Data acquisition** - cleaning (remove noise and inconsistent data), integration (combining possible multiple sources), selection (decide which data is relevant) and transformation (transformation, consolidation and aggregation);
- **Data mining** - essential process where intelligent methods are applied in order to extract data patterns;
- **Pattern evaluation** - identify truly interesting patterns representing knowledge;
- **knowledge presentation** - visualization and knowledge representation.

The use of this KDD process to build the user profiles and enrich the knowledge database is due to the huge volume of the input data for this solution, the CDR files. The KDD process is, already, a scientific area with a huge range of applications and successful implementations, allowing the development of automatic procedures and autonomous agent reasoning (programming).

An agent-based approach to the resolution of this problem is a major step through because:

- the ability of an agent-based system to develop autonomous tasks is, also, very well funded in the Multiagent Systems community, which contributes to the soundness of the system to develop;
- the huge amount of data (CDR) available is incompatible with a human-only approach, but, on the other hand, is the first requirement to the use of KDD processes.

Applying the KDD process to the proposed solution: in the data acquisition phase CDR files are read and processed: remove useless information (Figure 1-(a)), select the relevant information and make the necessary transformations(Figure 1-(b)); it's in the data mining phase that the profiles will be built: necessary to define the methods to extract the profiles(Figure 1-(c)); in the pattern evaluation phase, the profiles will be analyzed and compared, in order to identify patterns (Figure 1-(d)); at last, in the knowledge presentation phase, the gold is to enrich the knowledge database (Figure 1-(e)), providing methods for the telecommunication operator analysts to retrieve important and useful information.

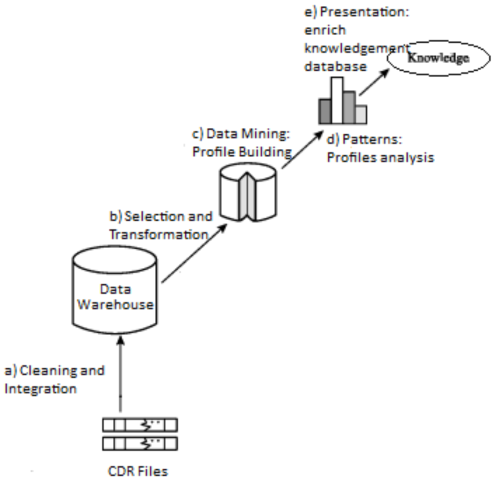
e) Presentation: enrich knowledgement database
Knowledge
c) Data Mining: Profile Building
d) Patterns: Profiles analysis
b) Selection and Transformation
Data Warehouse
a) Cleaning and Integration
CDR Files

Fig. 2. KDD process

4 Conclusion

MAS applications are ideal in order to build system designed to solve complex problems, which could not be solved by an individual agent. In a MAS application it is possible to solve a complex problem using different methods, skills and knowledge, distributing information and resources by the different agents. By supporting the solution proposal in a MAS model, it is possible to take advantage of all these facts giving it flexibility and extensibility properties.

Profiling is a technique that is used in several areas with good results. By applying this technique in the fraud problem, telecommunications operators will be able to cut down their losses and also get a better understanding of the problem. Dividing the profile in two, behavior and identity, allows serving two different purposes: identify possible fraudsters based on previous fraudsters profiles and to detect previous fraudsters that are using the network services again. Maintaining a profile history is also important in order to understand the evolution of a fraudster in the network.

The knowledge database, even if it has no direct impact in the detection of fraud, is very important for operators to better understand the problem they are facing, and optimize their resources (computer systems, analysts, thresholds) on detecting fraud.

As for the input data that will be used to build the profiles, the CDRs, it must be taken into account their huge number and, consequently, the size of the data that will input the methods and structures defined for profiling purposes. This was the main reason that led to the use of the KDD technique.

References

1. Communications fraud control association. http://http://www.cfca.org.
2. A. Bond and L. Grasser. Readings in distributed artificial intelligence. San Mateo, USA, 1988. Morgan Kaufmann Publishers.
3. I. V. Cadez, P. Smyth, and H. Mannila. Probabilistic modeling of transaction data with applications to profiling, visualization, and prediction. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 37–46, New York, NY, USA, 2001. ACM.
4. E. CORREIA, S. LUCAS, and A. LAMIA. Profiling: Uma técnica auxiliar de investigação criminal. *Análise Psicológica*, 2007.
5. K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, A. A. Nanavati, and A. Joshi. Social ties and their relevance to churn in mobile telecom networks. In *EDBT '08: Proceedings of the 11th international conference on Extending database technology*, pages 668–677, New York, NY, USA, 2008. ACM.
6. N. Du, B. Wu, X. Pei, B. Wang, and L. Xu. Community detection in large-scale social networks. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 16–25, New York, NY, USA, 2007. ACM.
7. G. McCrary. *Le profilage criminel à l'interieur et à l'extérieur du tribunal.* PUF, 2001.
8. J. McHugh, R. McLeod, and V. Nagaonkar. Passive network forensics: behavioural classification of network hosts based on connection patterns. *SIGOPS Oper. Syst. Rev.*, 42(3):99–111, 2008.
9. L. Montet. *Le profilage criminel.* PUF, 2002.
10. J. T. O'Brien. Telecommunications fraud. *The FBI Law Enforcement Bulletin*, 1998.
11. G. Olson, T. Malone, and J. Smith. Coordination theory and collaboration technology. Prentice Hall International Inc., 2001.
12. S. Russell and P. Norvig. Artificial intelligence, a modern approach. USA, 1995. Prentice Hall International Inc.
13. S. J. Stolfo, S. Hershkop, C.-W. Hu, W.-J. Li, O. Nimeskern, and K. Wang. Behavior-based modeling and its application to email analysis. *ACM Trans. Interet Technol.*, 6(2):187–221, 2006.
14. W.-G. Teng and M.-C. Chou. Mining communities of acquainted mobile users on call detail records. In *SAC '07: Proceedings of the 2007 ACM symposium on Applied computing*, pages 957–958, New York, NY, USA, 2007. ACM.

15. G. Weiss. Multiagent systems, a modern approach to distributed artificial intelligence. In *MIT Press*, USA, 1999.
16. M. Wooldrige. An introduction to multiagent systems. USA, 2002. John Wiley and Sons.
17. L. S. Wrighsman. *Forensic psychology.* Wadsworth, 2001.
18. K. Xu, Z.-L. Zhang, and S. Bhattacharyya. Internet traffic behavior profiling for network security monitoring. *IEEE/ACM Transactions on Networking*, 16(6):1241–1252, 2008.
19. Z. Zhu, W. Song, and J. Gu. A multi-agent and data mining model for tcm cases knowledge discovery. In *CCCM '08: Proceedings of the 2008 ISECS International Colloquium on Computing, Communication, Control, and Management*, pages 341–346, Washington, DC, USA, 2008. IEEE Computer Society.