EXTENDING THE RETSIM SIMULATOR FOR ESTIMATING THE COST OF FRAUD IN THE RETAIL STORE DOMAIN

Edgar Alonso Lopez-Rojas

Blekinge Institute of Technology
Department of Computer Science and Engineering

edgar.lopez@bth.se

ABSTRACT

RetSim is a multi-agent based simulator (MABS) calibrated with real transaction data from one of the largest shoe retailers in Scandinavia. RetSim allows us to generate synthetic transactional data that can be publicly shared and studied without leaking business sensitive information, and still preserve the important characteristics of the data.

In this paper we extended the fraud model of RetSim to cover more cases of internal fraud perpetrated by the staff and allow inventory control to flag even more suspicious activity. We also generated sufficient number of runs using a range of fraud parameters to cover a vast number of fraud scenarios that can be studied.

We then use RetSim to simulate some of the more common retail fraud scenarios to ascertain exactly the cost of fraud using different fraud parameters for each case.

Keywords: Multi-Agent Based Simulation, Retail Store, Fraud Detection,Retail Fraud, Synthetic Data.

1. INTRODUCTION

Fraud is an important problem in a number of different situations, and more specifically in retail stores is a very common problem in all countries. The economic impact for the loses can be substantial, this is why many major store retailers invest in security. However, how much to invest could be a political decision since many times it is hard to calculate the cost of possible fraudulent behaviour due to the multiple possible fraud causes. Once the fraud problem and cost of it is identified, it is easier for managers to take the prevention measures to lower the loses. For example, in one recent case the major US home improvement chain *Home Depot* was the target of a fraudulent return scam where two perpetrators netted several thousand dollars before being caught (FBI, 2013). Return fraud, i.e. the defrauding of a retail merchant by abusing the return process, alone is estimated to cost US retailers about 9 billion dollars yearly. To further illustrate the seriousness of the problem and try and combat it both EU and US recently started to mandate the use of fraud detection as one part of the minimum security requirements for financial services (ECB, 2013; Council, 2011).

Our approach to this problem makes use of the RetSim simulator as an strategy to measure and estimate the cost of loses and as a experimental laboratory for managers to apply measures to detect and discourage internal fraud. RetSim is a multi agent-based simulator (MABS) of a shoe store based on the transactional data of one of the largest retail shoe sellers in Sweden (Lopez-Rojas et al., 2013). RetSim uses this real data to develop and calibrate the model through statistical and a social Network Analysis (SNA) of the relations between staff and customers and generates a synthetic data set similar to the original one but without any unwanted disclosure of private information about either the staff or the customers.

Therefore, the aim of RetSim is the generation of synthetic data that can be used for fraud detection research and prognosis of fraud scenarios. With the RetSim simulator researchers and managers can test and measure in different fraud scenarios the cost and performance of fraud detection methods as simple as thresholds or even more elaborated such as triage models bases on thresholds, machine learning algorithms and others. The initial purpose of RetSim was to model a realistic data set that resembles the distribution of financial transactions found in an original given data set. Now that we have that we added fraudulent behaviour to study the cost of fraud in retail store.

In this paper we extended the fraud model of RetSim to cover more cases of internal fraud perpetrated by the staff and allow inventory control to flag even more suspicious activity. We also generated sufficient number of runs using a range of fraud parameters to cover a vast number of fraud scenarios that can be studied.

Our ultimate goal is for RetSim to be usable to model relevant scenarios to generate realistic data sets that can be used by academia, managers, security inspectors, students and others, to develop and reason about fraud detection methods without leaking any sensitive information about the underlying data. Synthetic data has the added benefit of being easier to acquire, faster and at less cost, for experimentation even for those that *have* access to their own data. We argue that RetSim generates data that usefully approximates the relevant aspects of the real data.

Outline: The rest of this paper is organised as follows: Section 2. introduce the topic of fraud detection for retail stores and present previous and related work on simulators. Section 3. describes the problem, which is the generation of synthetic data of a retail store system for estimating the cost of fraud. Sections 4. and 5. present our implementation and results of a MABS for our domain and shows the description of the extended fraud scenarios implemented on RetSim. We finish with a discussion and conclusions, including future work in section 6..

2. BACKGROUND AND RELATED WORK

Simulations in the domain of retail stores have traditionally been focused on finding answers to logistics problems such as inventory management, supply management, staff scheduling and customer queue reductions (Chaczko and Chiu, 2008; Schwaiger and Stahmer, 2003; Bovinet, 1993). We find no research focusing on simulations generating fraud data to be used for fraud detection in retail stores. Therefore, we recently introduced RetSim with the purpose of fraud detection research. In this article we built upon previous version of RetSim to study the cost of specific fraud scenarios, including agents using known fraud behaviour patterns (Lopez-Rojas et al., 2013).

Anonymization techniques have been used to preserve the privacy of sensitive information present in data sets. But de-anonymizing data sets is not an insurmountable task, far from it (Narayanan and Shmatikov, 2009). For this reason we have decided to use simulation techniques to keep specific properties of the original data set, such as statistical and social network properties, and at the same time providing an extra layer of insulation that pure anonymization does not provide.

There are tools such as IDSG (IDAS Data and Scenario Generator (Lin et al., 2006)) that were developed for the purpose of generating synthetic data based on the relationship between attributes and their statistical distributions. IDSG was created to support data mining systems during the testing phase, and it has been used to test fraud detection systems. Our approach differs in that we are implementing an agent-based model which is based on agent micro behaviour rather, than a fixed statistical distribution of macro parameters.

With the current popularity of social networks, such as *Facebook*, the topic of Social Network Analysis (SNA) has seen interest in the research community (Alam and Geller, 2012). Social Network Analysis is currently being combined with *Social Simulation*. Both topics support each other in the representation of interactions and behaviour of agents in the specific context of social networks. However, there is no work addressing the question of customer/salesman-interaction, that we are aware of.

Other methods to generate the necessary fraud data have been proposed by (Yannikos et al., 2010; Lundin, 2002; Kargupta et al., 2003; Evfimievski et al., 2002; Agrawal and Aggarwal, 2001). The work by (Yannikos et al., 2010) lets the user specify the assumptions about the environment at hand; i.e., there is no need for access to real data. However, this will certainly affect the quality of the synthetic data. The work by (Lundin, 2002) makes use of a small sample of real data to generate synthetic data. This approach is similar to ours. However, the direct use of real data to prime the generation of synthetic data is limited in that it makes it harder to generate realistic data with other characteristics that those of the original real data (Yannikos et al., 2010). The work by (Kargupta et al., 2003) focused on privacy-preserving methods for data mining. However, that method also does not have the possibility of generating realistic data with other characteristics than those of the original data. In our work, we use social simulation, which makes it possible to change the parameters of the agents in the model to create realistic synthetic data, potentially producing emergent behaviour in the logs which is hard to produce in other ways.

3. PROBLEM

The RetSim simulator was previously used to solve the problem of finding a synthetic data set that realistically represent a given real data set without disclosing any specific information about customers or staff members. Now with this paper we aim to solve the question of how much are the loses due to fraud giving a specific scenario of fraud in a retail store. This problem is of particular interest for managers of stores, mainly because the investment on security is limited, therefore the impact of each of these fraud models and expected fraud scenarios will give an idea to a manager of the cost of fraud. Historically on a real store the cost is only estimated but can not be exactly calculated due to the disguise nature of the fraud.

By using simulators, we can be certain of the profit of each fraudulent agent over the time. We generate synthetic data that contains flagged fraud behaviour, therefore it simplifies the process of estimating the cost of loses due to specific fraud. These estimations can be used to take informed decisions because they are a good approximation of a real scenario.

4. MODEL AND METHOD

RetSim uses the MABS toolkit MASON version 17 which is implemented in Java (Luke, 2005). We selected MASON because it is: multi-platform, supports parallelisation, and fast execution speed in comparison with other agent frameworks. This is especially important for multiple running and computationally expensive simulations such as RetSim (Railsback et al., 2006).

Our model contains the following entities and behaviours. The *Store* is the main entity of the simulation, it contains all the variables and states required to run the simulation such as: *Salesmen*, *Customers*, *Products*, *Frequencies* and other parameters used to calibrate the model.

During the initialization of the simulation a store load all products, set up an inventory and creates the salesmen

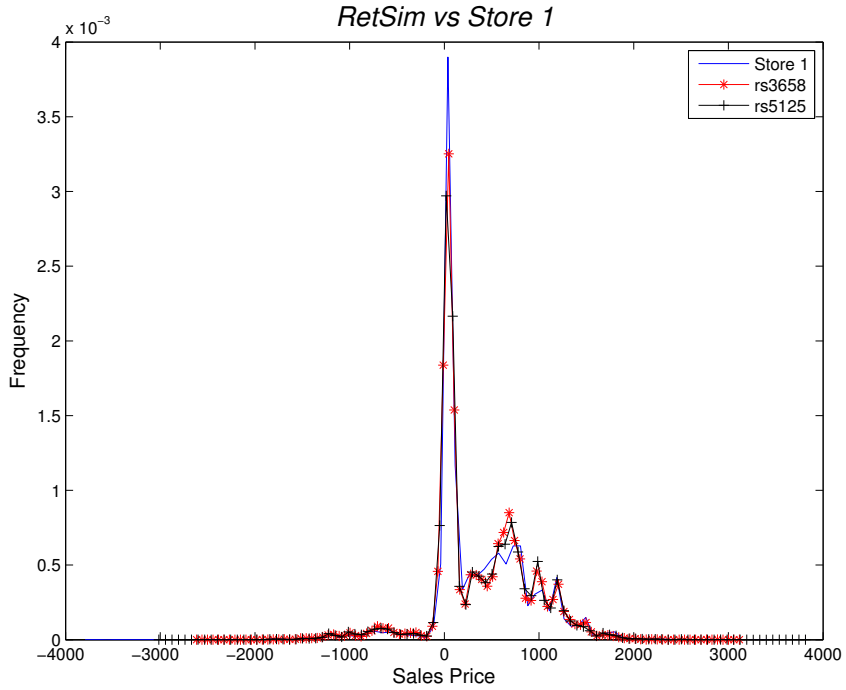

Figure 1: Overlap of Two Runs of RetSim vs Real Data

according to the parameters of the store. We calibrated RetSim to consider each day as a step in the simulation. During a single step of the simulation some of the salesmen became active and others inactive or not working that day. The salesmen are distributed in a virtual geographical space on the store. Once the day starts an average of customers are instantiated and distributed around the geographical space.

The *basic principle* of this model is the concept of a commercial transaction. The process of creating financial transactions starts when a salesman sense nearby customers in the *need-help* state and offers help. There are two different outcomes: either a transaction takes place or not, similar to the real world.

There are no known instances of fraud in the real data (as certified by the data owner). So we modelled malicious agents with fraud behaviour, by programming agents that behave according to some known or hypothesised retail fraud case.

The following retail fraud scenarios are based on selected cases from the Grant Thornton report (Member and Council, 2009). As can be seen below, we extended the RetSim model by adding different fraud scenarios. The initial purpose of RetSim was to model a realistic data set that resembles the distribution of financial transactions found in an original given data set. Now that we have that we added fraudulent behaviour to study the cost of fraud in retail store.

The *Refunds fraud scenario* is perhaps the most common cause of fraud and includes cases where a salesman creates fraudulent refund slips, keeping the cash refund for him- or herself.

In terms of the object model used in RetSim, the refund scenario was simulated by estimating the average number of refunds per sale and the corresponding standard deviation. We used these statistics to simulate refunds in the RetSim model. Fraudulent salesmen will perform normal refunds, as well as fraudulent ones. The volume of fraudulent refunds was modelled using a salesman specific parameter. The ``red flag" for detection would in this case be a high number of refunds for a salesman.

The *Coupon reductions/discounts fraud scenario* is perhaps the second responsible for loses and includes cases where the salesman registers a discount on the sale without telling the customer; i.e., the customer pays the full sales price, and the salesman keeps the difference.

In terms of the object model used in RetSim, the coupon reduction/discounts scenario was implemented by estimating the average number of cancellations per sale and the corresponding standard deviation. Using these statistics we simulated discounts in the RetSim model. Fraudulent salesmen performed normal discounts, as well as fraudulent ones. The volume of fraudulent discounts was modelled using a salesman-specific parameter. The ``red flag" for detection would in this case be a high number of discounts for a salesman with a relatively low number of average sales.

We extended the original RetSim by adding inventory control over products. Our initial assumptions were that the replenish of the inventory was performed automatically and therefore we did not focus on this aspect. But in reality inventory control is one of the more effective ways to detect fraud. Unfortunately performing inventory control is an expensive procedure in most of the stores due to

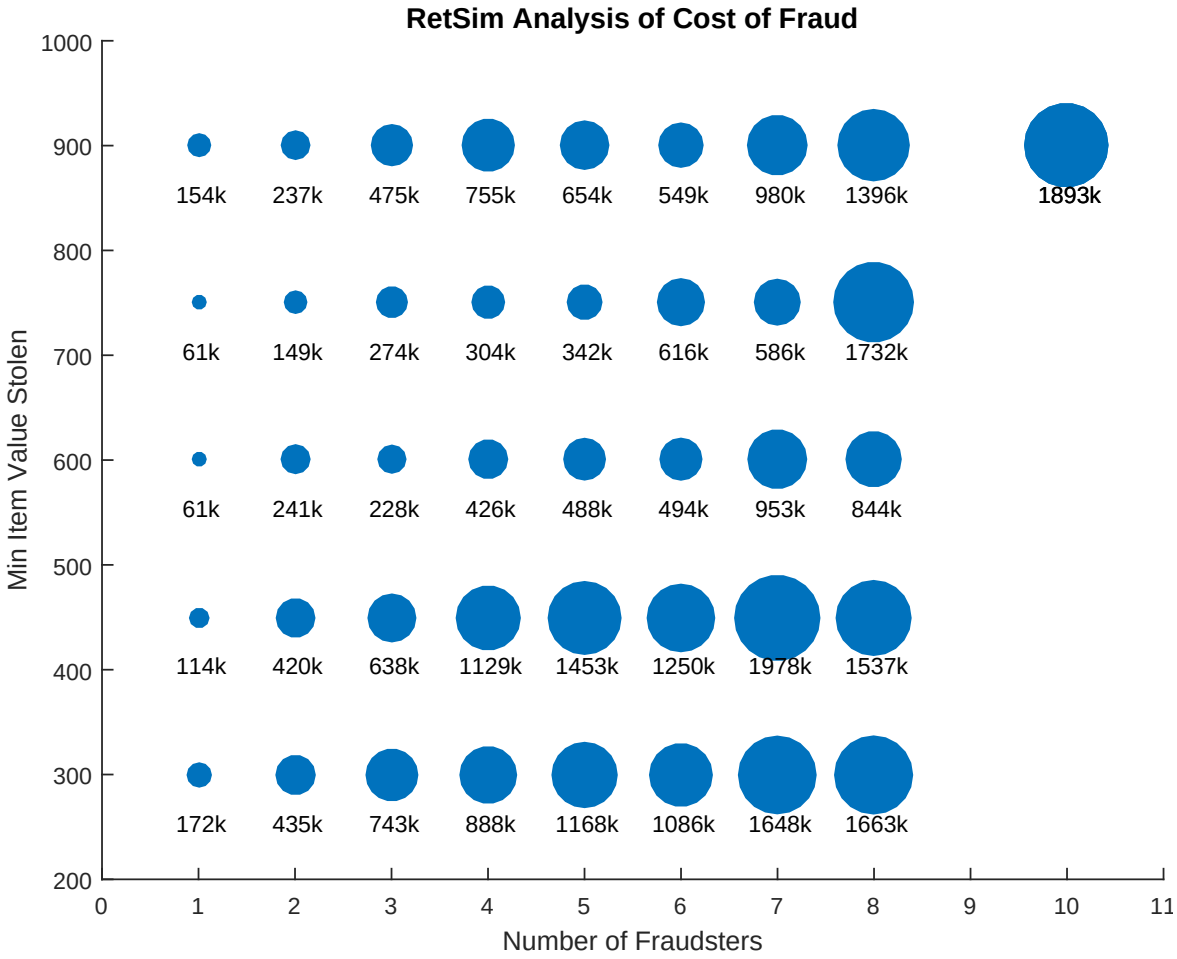

Figure 2: Analysis of Cost of Fraud

manual counting.

There are other possible scenarios, but as mentioned in the introduction return fraud (both by customers and sales staff alike) is a major problem, so we have chosen to focus on return fraud and the structurally similar discount fraud, as these are common and serious.

5. RESULTS

We built upon previous successful simulations that were calibrated from a Store taken from the original data set of our retailer. Figure 1 shows the results from a comparison of sales frequency for items of 2 runs of RetSim versus the original store (named Store 1). We can clearly see that our simulator successfully models the sales of Store 1. We can analyse that items that cost less are the most wanted products by the customers. For convenience and protection of disclosure of the business we decided to use a fictitious monetary unit that we will name just *units*.

From an analysis of figure 1, a manager can design experiments to study the phenomenon of fraud performed by the staff in the previous scenarios explained (Refunds fraud and discounts fraud).

Our experiment aims to calculate the total cost of loses if one or many employees are responsible for fraud. One common behaviour is of fraudsters is to mentally set some limits about the min or max value of the item stolen. Many fraud controls nowadays are performed using simple thresholds around the prices. One way to avoid these controls is by carefully selecting items with value below. We will assume that the fraudster mentally set up a minimum value for stealing to make it worth the risk of being caught by the manager.

We setup our experiment to run 100 times under a specific setting. We iterated around minimum values starting from 300 *units* up to 900 *units*. We use a probability of 10% of the time that a salesmen should do the checks for performing fraud. In total we collected 500 runs of RetSim and summarize our cost analysis in figure 2.

Figure 2 shows one way to visualize the total cost of fraud performed by a different number of staff members. This figure allows a manager to study better the phenomenon of fraud in a store. One can notice for instance that the most of the items sold are around the 450 *units* price. So the profit of a fraudster that steals around this minimum value will be in many cases higher than other that decide to steal only high value articles. This information is particularly important when we require to implement controls over the refunds of items that cost more or less than 450 *units*. We can also see that below this value the amount of loses is considerable when then number of

staff members involved in fraud increases.

Besides the controls and thresholds that can be used for fraud, a major contribution of the RetSim to the managers is to measure the total cost of fraud. With this information a manager can decide how much to invest in any required fraud detection control. In many cases it is *"acceptable"* for managers to deal with low loses if the investment in fraud is higher than the fraud that can be prevented.

6. CONCLUSIONS

RetSim is a simulator of a retail store that generates transaction data set of diverse fraud scenarios, in this paper we extend the fraud model and measure the cost of total loses of each scenario. The cost of fraud in different scenarios can be estimated by summing the profit from each malicious agent (usually known as fraudsters). These estimations can be used to take informed decisions about how much should be invested in fraud controls, because they are a good approximation of real fraud scenarios.

Synthetic data sets generated with RetSim can aid academia, managers of companies and governmental agencies in testing their methods, in exploring the cost of different fraud scenarios and the performance of different fraud detection methods while maintaining similar conditions by using the same test data set, or in generally reasoning about the limits of effectiveness of fraud detection.

Future work on RetSim may include the addition of more relevant fraud models as well as tools for setting up experiments with higher degree of parametrization that includes more variables such as variations (increase or decrease) of sales with respect to the original store.

ACKNOWLEDGMENTS

This work is part of the research project "Scalable resource-efficient systems for big data analytics" funded by the Knowledge Foundation (grant: 20140032) in Sweden. The author thanks Dr. Stefan Axelsson, main supervisor of doctoral studies, for encouraging this solo research as a preparation for upcoming doctor degree next year.

REFERENCES

Dakshi Agrawal and CC Aggarwal. On the design and quantification of privacy preserving data mining algorithms. *PODS '01 Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2001.

SJ Alam and Armando Geller. Networks in agent-based social simulation. *Agent-based models of geographical systems*, pages 77--79, 2012.

JW Bovinet. *RETSIM: A Retail Simulation with a Small Business Perspective*. West Pub. Co., Minneapolis/St. Paul, 1993. ISBN 0314016708.

Z. Chaczko and C.C. Chiu. A smart-shop system - Multiagent simulation system for monitoring retail activities. *20th European Modelling and Simulation Symposium*, pages 20--26, 2008.

FFIE Council. Supplement to Authentication in an Internet Banking Environment. *URl: http://www. ffiec. gov/pdf/Auth-ITS-Final*, pages 206--222, 2011.

(European Central Bank) ECB. Recommendations for the Security of Internet Payments. Technical Report January, 2013.

Alexandre Evfimievski, Ramakrishnan Srikant, Rakesh Agrawal, and Johannes Gehrke. Privacy preserving mining of association rules. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02*, page 217, 2002. doi: 10.1145/775079.775080.

FBI. Ticket Switch Fraud Scheme at Home Deopt, 2013. URL `http://www.fbi.gov/buffalo/press-releases/2013/mexican-man-charged-in-ticket-switch-fraud/-scheme-at-home-depot`.

H. Kargupta, S. Datta, and Q. Wang. On the privacy preserving properties of random data perturbation techniques. *Third IEEE International Conference on Data Mining*, pages 99--106, 2003. doi: 10.1109/ICDM.2003.1250908.

P.J. Lin, B. Samadi, and Alan Cipolone. Development of a synthetic data set generator for building and testing information discovery systems. In *ITNG 2006.*, pages 707--712. IEEE, 2006. ISBN 0769524974.

Edgar Alonso Lopez-Rojas, Stefan Axelsson, and Dan Gorton. RetSim: A Shoe Store Agent-Based Simulation for Fraud Detection. *The 25th European Modeling and Simulation Symposium*, 2013.

S. Luke. MASON: A Multiagent Simulation Environment. *Simulation*, 81(7):517--527, July 2005. ISSN 0037-5497. doi: 10.1177/0037549705058073.

Emilie Lundin. A synthetic fraud data generation methodology. In *Information and Communications Security*, pages 265--277. Springer, Berlin Heidelberg, 2002.

Associate Member and Advisory Council. Reviving retail Strategies for growth in 2009 Executive summary, 2009. URL `http://www.grantthornton.com/staticfiles/GTCom/files/Industries/Consumer&industrialproducts/Whitepapers/Revivingretail_Strategiesforgrowthin2009.pdf`.

Arvind Narayanan and Vitaly Shmatikov. Deanonymizing Social Networks. *2009 30th IEEE Symposium on Security and Privacy*, pages 173--187, May 2009. doi: 10.1109/SP.2009.22.

S. F. Railsback, S. L. Lytinen, and S. K. Jackson. Agent-based Simulation Platforms: Review and Development Recommendations. *Simulation*, 82(9):609--623, September 2006. ISSN 0037-5497. doi: 10.1177/0037549706073695.

Arndt Schwaiger and B Stahmer. SimMarket: Multiagent-based customer simulation and decision support for category management. *Multiagent System Technologies*, pages 74--84, 2003.

York Yannikos, Frederik Franke, Christian Winter, and Markus Schneider. 3LSPG : Forensic Tool Evaluation by Three Layer Stochastic Process-Based Generation of Data. In *4th International Workshop in Computational Forensics*, pages 200--211, Tokyo, Japan, 2010.

AUTHORS BIOGRAPHY

MSc. Edgar A. Lopez-Rojas

Edgar Lopez is a PhD student in Computer Science and his research area is Multi-Agent Based Simulation, Machine Learning techniques with applied Visualization for fraud detection and Anti Money Laundering (AML) in the domains of retail stores, payment systems and financial transactions. He obtained a Bachelors degree in Computer Science from EAFIT University in Colombia (2004). After that he worked for 5 more years at EAFIT University as a System Analysis and Developer and partially as a lecturer. He obtained a Masters degree in Computer Science from Linköping University in Sweden in 2011 and a licentiate degree in computer science (a degree halfway between a Master's degree and a PhD) in 2014.