RETSIM: A SHOE STORE AGENT-BASED SIMULATION FOR FRAUD DETECTION

Edgar Alonso Lopez-Rojas[(a)], **Stefan Axelsson**[(b)], and **Dan Gorton**[(c)]

[(a),(b)]Blekinge Institute of Technology , School of Computing
[(c)]KTH Royal Institute of Technology , Department of Transport Science , Center for Safety Research

[(a)] edgar.lopez@bth.se, [(b)] stefan.axelsson@bth.se, [(c)] dan.gorton@abe.kth.se

ABSTRACT

RetSim is an agent-based simulator of a shoe store based on the transactional data of one of the largest retail shoe sellers in Sweden. The aim of RetSim is the generation of synthetic data that can be used for fraud detection research. Statistical and a Social Network Analysis (SNA) of relations between staff and customers was used to develop and calibrate the model. Our ultimate goal is for RetSim to be usable to model relevant scenarios to generate realistic data sets that can be used by academia, 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.

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

1. INTRODUCTION

In this paper we introduce *RetSim*, a **Ret**ail shoe store **Sim**ulation, built on the concept of Multi Agent-Based Simulation (MABS). RetSim is based on the historical transaction data provided by one of the largest Nordic shoe retailers. This data contains several hundred million records of diverse transactional data from a few years ago, and covering several years. That is, this data is recent enough to reflect current conditions, but old enough to not pose a risk from a competitor analysis standpoint.

The defence against fraud is an important topic that has seen some study. In the retail store the cost of fraud are of course ultimately transferred to the consumer, and finally impacts the overall economy. Our aim with RetSim is to learn the relevant parameters that governs the behaviour in a retail store to simulate *normal* behaviour, which is our focus in this paper.

The main contribution and focus of this paper is a method to generate anonymous synthetic data of a retail store, that can then be used as part of the necessary data for the development of fraud detection techniques. Even so, the data set generated could also be the basis for research in other fields, such as demand prediction, logistics and demand/supply research.

Later we plan to address the actual fraud and develop techniques to develop malicious agents to inject fraudulent and anomalous behaviour, and then develop and test different strategies for detecting these instances of fraud. Even though we do not address these issues in this paper, we describe some typical scenarios of fraud in a retail store. As this is our ultimate goal, fraud heavily influenced the design of RetSim.

The main goal of developing this simulation is that it enables us to share realistic fraud data, without exposing potentially business or personally sensitive information about the actual source. As data relevant for computer security research often is sensitive due to a multitude of reasons, i.e. financial, privacy related, legal, contractual and other, research has historically been hampered by a lack of publicly available relevant data sets. Our aim with this work is to address that situation. However, simulation also have other benefits, it can be much faster and less expensive than trying different scenarios of fraud, detection algorithms, and personnel and security policy approaches in an actual store. The latter also risks incurring e.g. unhappiness amongst the staff, due to trying e.g. an ill advised policy, which leads to even greater expense and unwanted problems.

Outline: The rest of this paper is organized as follows: Section 2. introduce the topic of fraud detection for retail stores and present related work. Sections 3. describes the problem, which is the generation of synthetic data of a retail store. Section 4. shows a data analysis of the current data. Section 5. presents an implementation of a MABS for our domain and shows the description of some retail fraud scenarios. We present our results and verification of the simulation in section 7. and finish with a discussion and conclusions, including future work in section 8..

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 for customer queue reductions (Chaczko and Chiu, 2008; Schwaiger and Stahmer, 2003; Bovinet, 1993).

There is currently a lack of research in the area of simulation of the retail environment for fraud detection and here is where we focus in this work.

We have previously analysed the implications of using machine learning techniques for fraud detection using a synthetic dataset (Lopez-Rojas and Axelsson, 2012a). We then built a simple simulation of a financial transaction system based on these assumptions, in order to overcome our limitations and lack of real data (Lopez-Rojas and Axelsson, 2012b). However, this work was not based on any underlying data, but rather on assumptions of what such data could contain. Here we continue and build a realistic simulation based on a real data set that in the future can be used to test diverse fraud detection techniques.

Data mining based methods have been used to detect fraud (Phua et al., 2010). This lead to the result that machine learning algorithms can identify novel methods of fraud by detecting those transactions that are different (anomalous) in comparison with the benign transactions. This problem in machine learning is known as novelty detection. Supervised learning algorithms have previously been used on a synthetic data set to prove the performance of outliers detection (Abe et al., 2006), however this has not been done over transactional data. There are tools such as IDSG (IDAS Data and Scenario Generator (Lin et al., 2006)) which was developed with 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 their test phase and it has been used to test fraud detection systems.

Nowadays with the popularity of social networks, such as *Facebook*, the topic of Social Network Analysis (SNA) has been given special interest in the research community (Alam and Geller, 2012). Social Network Analysis is a topic that is currently being combined with Social Simulation. Both topics support each other for the benefit of representing the interactions and behaviour of agents in the specific context of social networks.

Our approach aims to fill the gap between existing methods and provide researchers with a tool that generates reliable data to experiment with different fraud detection techniques and compare them with other approaches.

3. PROBLEM

Fraud and fraud detection is an important problem that has a number of applications in diverse domains. However, in order to investigate, develop, test and improve fraud detection techniques one needs detailed information about the domain and its specific problems.

There is a lack of data sets available for research in fields such as money laundering, financial fraud and illegal payments. Disclosure of personal or private information is only one of the many concerns that those that own relevant data have. This leads to in-house solutions that are not shared with the research community and hence there can be no mutual benefit from free exchange of ideas between the many worlds of the data owners and the research community.

After describing the problem we formulated the main research question that we address on this paper:

RQ *How could we model and simulate a retail shoe store and obtaining a realistic synthetic data set for the purpose of fraud detection?*

4. Data Analysis

To better understand the problem domain we began by performing a data analysis over the historical data provided by the retailer. We are interested in finding the necessary and sufficient attributes to enable us to simulate a realistic scenario in which we could reason about and detect interesting cases of fraud.

We initially started by selecting five stores that represent different sizes of store in the company. We selected two big stores, one medium and two small. We extracted statistical information from the data set, presented in table 1. All prices given are in a fictitious currency.

Due to a lack of space we will focus our presentation of the analysis on one of the big stores by sales volume, store one. Store one is relatively richer in data than the smaller stores. This is specially interesting, since we are more likely to find actual cases of fraud in a big store. We took a sample that comprises the sales during a year. We selected the transaction tables that detail cash flow and the articles inventory, which give us a good idea of how many transactions a big store can produce in a year, and how many different types of articles and their quantities that are sold in a year.

Table 1: Statistical analysis of five stores during one year

Stat-Store	1	2	3	4	5
Transactions	147037	180626	44446	37776	28456
Receipts	43406	38376	10094	8595	7619
Returns	9,25%	9,67%	11,43%	9,89%	9,33%
Members	5509	6381	1375	1152	16
Mem. Rec	16,02%	14,14%	18.12%	22,33%	0,56%
Avg. Price	762,49	772,32	665,2	575,93	409,62
Std. Price	494,52	514,51	459,05	616,74	416,36

4.1. Statistical Analysis

The store one sample contains 147 037 records of transactions. Note that this does not mean receipts, as a single receipt can produce several records. The retailer runs a fidelity program that allows customers to register their purchases. From this one store we identified 5509 unique members that made at least one purchase during the period resulting in 16,02% of the receipts. This means that the majority of receipts belongs to unidentified customers. However in all these records we can identify the item(s), sales price and the sales clerk.

We extracted statistical information, presented in table 1 and plotted in figure 1 which represents the sales summary per day and figure 2 which shows the number of customers per day.

Some observations that stand out in the data set:

- There were 67 receipts where the customer did not pay anything for the item, it means that the discount was 100% without returning any other article to the store. This could possible be due to a fraud, and when investigated could be used for injecting malicious behaviour.

- It was very rare for a customer to buy the same article more than once in the same purchase, this happened only three times during the year.

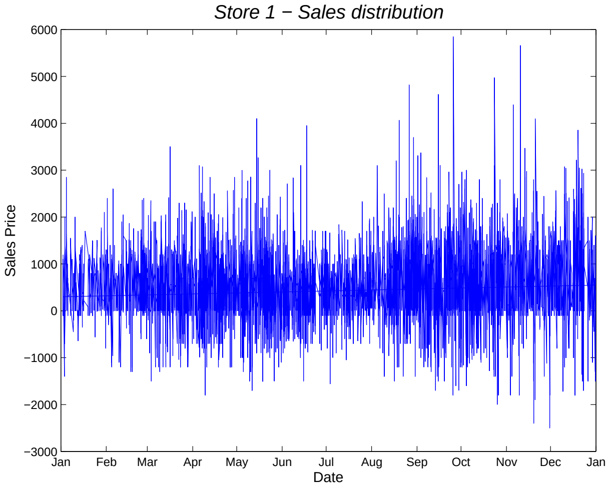

Figure 1: Store one - sales distribution

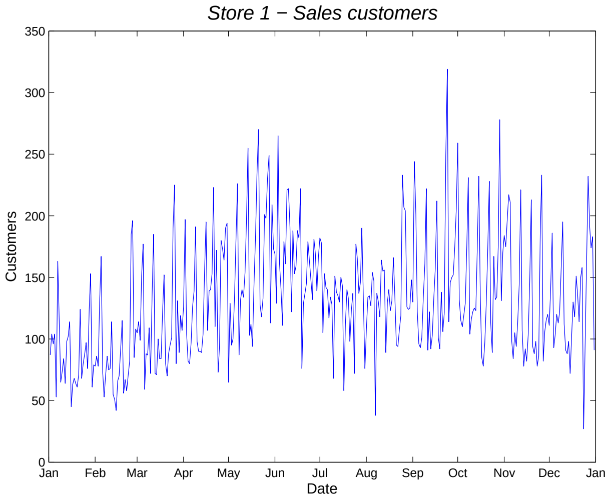

Figure 2: Store one - number of customers per day

We then investigated the performance of the staff. We divided the sales staff into three categories: *top*, *medium* and *low*. *Top* refers to staff that works regularly at the store. *Medium* refers to seasonal staff that works usually for a period between one and three months. Finally *Low* refers to staff that worked for less than one month. Table 2 shows the distribution of frequencies found in the data. Top sale clerks work an average of 66% of the time at the store, and they are only 22% of the total number of sales staff.

Table 2: Sales clerk frequency

Type	Avg. Days	Avg. Cust	Std. Cust	Quantiy
Top	155,75	45,43	28,17	22,22%
Med	63,20	38,97	23,83	11,11%
Low	13,57	33,93	16,68	66,67%

Table 3: Article categories

Category	Probability	Rank
Top	0,2705	+1000
High	0,2122	100-999
Medium	0,1109	20-99
Low	0,3495	3-19
Unfreq	0,0569	1-2

4.2. Network Analysis

Fraud has traditionally had a strong association to network analysis. Due to the possibility of several actors participating in a specific fraud in order to confuse the investigators and dilute the evidence. Another advantage of a network analysis is the ability to visualize the network by using different layout algorithms such as *Force Atlas* or *Yifan Hu* (Hu, 2005). In this project we used the *Gephi* software, that does network analysis and allows the use of different layout algorithms for the visualization of the network (Bastian et al., 2009).

We can create a network based on the interactions between each of the sales clerks and their respective customers. For the weight of the edges we use the total sales price with respect to each customer. Figure 3 shows one way to visualize the sample data extracted from the database using *Yifan Hu* layout.

The network topology resembles a hub topology, where the sales clerks are the central nodes of the hubs, and a few customers that have been helped by more than one sales clerk act as bridges between the hubs.

The store one sample contains 5545 nodes where 36 of them are sales staff, with the rest being customers. The network contains 6120 edges that connects the sales staff and customers. Each edge weight represents the total amount of purchases per customer. Table 4 show more information about the network used for calibrating the simulation.

Figure 3 shows a visualization of the network for the store, the size of the nodes is determined by the outdegree of the sales clerks. The number inside the nodes also represent the number of customers that were helped

Table 4: Network Analysis

Statistic	Store one
Nodes	5.545
Sales Clerks	36
Customers	5.509
Avg. Degree	1.104
Diameter Undirected	10
Avg. Path Undirected	3.98

by the sales clerk. The In-degree distribution can be better visualized in figure 4.

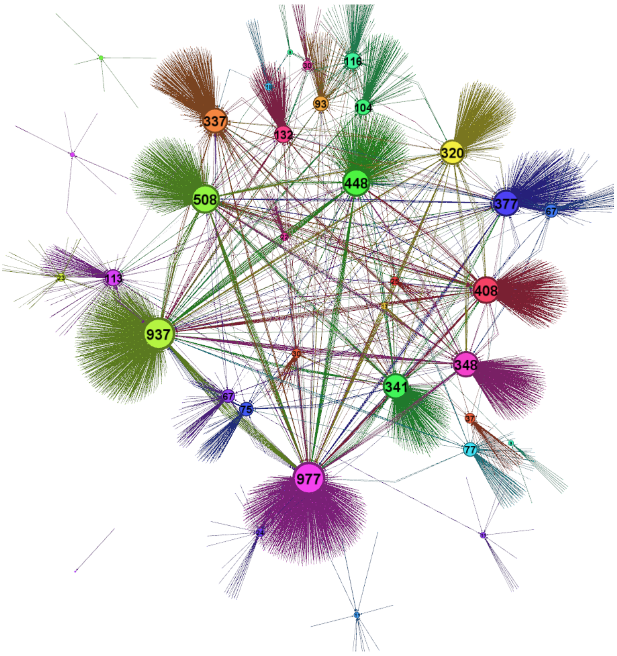

Figure 3: Store one - Network of customers and sales clerks

From the network analysis there is a lot of data we can use for our model, e.g. that 90.26% of the members have been helped by only one sales clerk, as described by the out-degree distribution.

5. MODEL AND METHOD

The design of RetSim was based on the ODD model introduced by Grimm et al. (2006). ODD contains 3 main parts: *Overview*, *Design Concepts* and *Details*.

5.1. Overview

5.1.1. Purpose

We aim to produce a simulation that resembles a real retail store. Our main purpose is to generate a synthetic data set of business transactions that can be used for the development and testing of different fraud detection techniques. It is important due to the difficulty to find diverse and enough cases of fraud in a real data set. However this is not the case of a simulated environment, where fraud can be injected following known patterns of fraud.

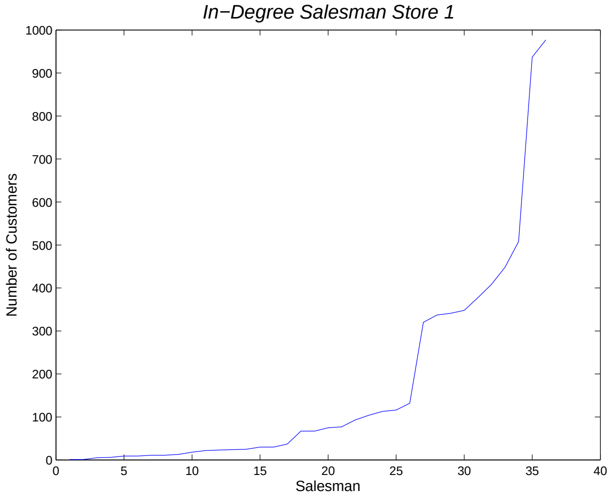

Figure 4: Store one - Customers per sales clerks

5.1.2. Entities, state variables and scales

There are three agents in this simulation: *Manager*, *Sales clerk* and *Customer*.

Manager This agent decides the price, check inventory and order new items.

Sales clerk Is in charge of promoting the items and issues the receipt after each sale. A sales clerk can be in state busy when the clerk is serving its maximum amount of customers.

Customer The behaviour is determined by the goal of purchasing one or several items. A customer is in an active *need-help* state, when no sales clerk is assisting with shopping.

5.1.3. Process overview and scheduling

During a normal step of the simulation a customer enters the simulation, and a sales clerk sense nearby customers in the *need-help* state and offers help. There are two different outcomes: Either a transaction takes place, with probability p, or no transaction takes place with, trivially, probability $1 - p$.

The time granularity of the simulation is that each step represents a day of sales. So a normal week has seven steps and a month will consist of around 30 steps. We do not make any explicit distinction between specific days of the week. Instead we handle differences between days by using a different distribution of the customers per day (see figure 2).

5.2. Design Concepts

The *basic principle* of this model is the concept of a commercial transactions. We can observe an *emergent* social network from the relation between the customers and the sales clerks. Each of the customers have the *objective* of purchasing articles from the store. The sales clerks *objective* is to aid the customers and produce the receipt necessary for the generation of the data set. Managers play a special role in the simulation. They serve as the schedulers for the next step of the simulation. Given the specific step of the simulation the manager generate a supply of customers for the next day and activate or deactivate specific sales clerks in the store. In our virtual environment the *interaction* between agents is always between sales clerk and customer. Purchase articles from another customer or selling articles to a sales clerk is not permitted.

Customers and sales clerks can scout the store in any radial direction from their current position and search or offer help, respectively.

The agents do not perform any specific learning activities. Their behaviour is given by probabilistic Markov models where the probabilities are extracted from the real data set.

5.3. Details

5.3.1. Initialization

The simulation starts with a number of sales clerks that serve the customers, an initial number of customers and one manager that does the scheduling.

The In-degree distribution is used as an indication of how good a sales clerk can be. Each sales clerk is assigned an in-degree value in each step of the simulation when the sales clerk searches for customers in need of assistance. The bigger their in-degree the more customers they can help.

5.3.2. Input Data

RetSim has different inputs needed in order to run a simulation. The input data concerns the distributions of probabilities for scheduling the sales clerks, the items that can be purchased and different statistic measures for the customers. A CSV file which contains an identifier, description, price, quantity sold and total sales specify these inputs. For setting the parameters, including the name of the CSV-file, we use a parameter file that is loaded as the simulation starts or the can also be set manually in the GUI.

5.3.3. Submodels

Figure 5 shows the different use cases of the agents. This model represent the different actions that an agent can take inside the system.

Manager scheduler This agent is in charge of scheduling the next step of the simulation. There is only one manager per store. This agent creates the new customers that are going to arrive to the store according to a distribution function extracted from the original data set. The manager also allocate the sales clerks that are going to be active during the this step of the simulation.

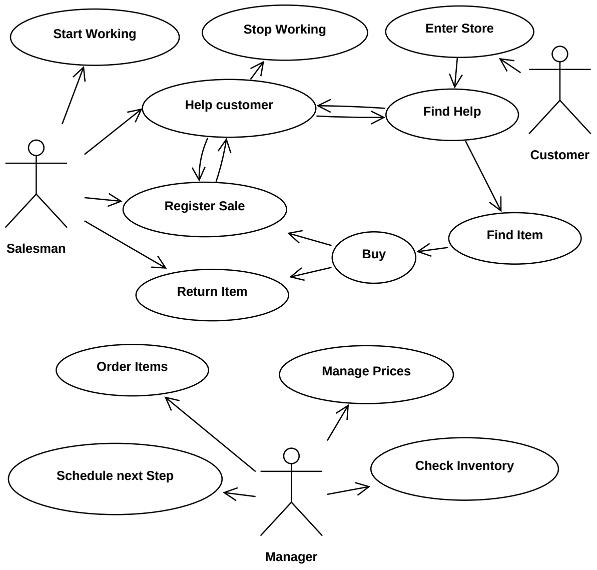

Figure 5: RetSim Use Case Diagram

Customer finder Is performed by the sales clerk and it starts with the agent searching nearby for a customer that is not being helped by an other sales clerk. Once the contact is established a sale is likely to occur with a certain probability.

Sales clerk finder Customers that are still in need for help can also look for nearby sales clerks. This again could lead to a sale.

Network generation Every time a transaction is performed between a customer and a sales clerk, an edge is created in the network composed of the customers and the sales clerks in attendance. The weight of the edge represent the sales price. The network grows by the inclusion of new customers or sales clerks.

Item selection for purchasing Items are classified into 5 different categories according to their quantity or units sold (see table 3). From the original data we extracted the probabilities of each of the categories and quantities. A customer can also purchase more than one item.

Item return after purchasing A customer can also decide to return a purchased item with a certain probability p.

Log of receipt transactions Each time an item is purchased a receipt is created. A receipt contains the information about the customer, sales clerk, item(s), quantities, sales price, date and discount if any.

6. Fraud Scenarios in a Retail Store

In this section we describe how three examples of retail fraud can be implemented in RetSim. These fraud scenarios are based on selected cases from Thornton (2009) report. As can be seen in section 5., the different scenarios can be implemented in almost the same way. Furthermore, a fraudulent sales clerk will probably use several different methods of fraud, which means that RetSim needs to be able to model combinations of all fraud scenarios implemented. Although the implementation of these scenarios are out of the scope of this paper, we include a description and explain how to implement them in RetSim.

6.1. Sales cancellations

This scenario includes cases where the sales clerk cancels some of the items in the sale without telling the customer, i.e., the customer pays the full sales price, and the sales clerk keeps the difference. In terms of the object model used in RetSim the sales cancellation scenario can be implemented by the following setting: Estimate the average number of cancellations per sale and the corresponding standard deviation. Use these statistics for simulating normal cancellations in the RetSim model. Fraudulent sales clerks will perform normal cancellations, as well as fraudulent once. The volume of fraudulent cancellations can be modelled using a sales clerk specific parameter. The "red flag” for detection will in this case be a high number of cancellations for a sales clerk with a low number of average sales.

6.2. Refunds

This scenario includes cases where the sales clerk creates fraudulent refund slips, keeping the cash refund for himor herself. In terms of the object model used in RetSim the refund scenario can be implemented by the following setting: Estimate the average number of refunds per sale and the corresponding standard deviation. Use these statistics for simulating refunds in the RetSim model. Fraudulent sales clerks will perform normal refunds, as well as fraudulent once. The volume of fraudulent refunds can be modelled using a sales clerk specific parameter. The "red flag” for detection will in this case be a high number of refunds for a sales clerk.

6.3. Coupon reductions/discounts

This scenario includes cases where the sales clerk registers a discount on the sale without telling the customer, i.e., the customer pays the full sales price, and the sales clerk keeps the difference. In terms of the object model used in RetSim the coupon reduction/discounts scenario can be implemented by the following setting: Estimate the average number of cancellations per sale and the corresponding standard deviation. Use these statistics for simulating discounts in the RetSim model. Fraudulent sales clerks will perform normal discounts, as well as fraudulent ones. The volume of fraudulent discounts can be modelled using a sales clerk specific parameter. The "red flag" for detection will in this case be a high number of discounts for a sales clerk with a low number of average sales.

7. RESULTS

RetSim uses the Multi-Agent Based Simulation toolkit MASON which is implemented in Java (Luke, 2005). MASON offers several tools that aid the development of a MABS. We justified our choice mainly for the benefits of supporting multi-platform, parallellization, good execution speed in comparison with other agent frameworks; which is specially important for computationally intensive simulations such as RetSim (Railsback et al., 2006). RetSim can be run with GUI, that helps the user see the states and relations between the sales clerks (bigger circles) and customers, as can be seen in the example in figure 6.

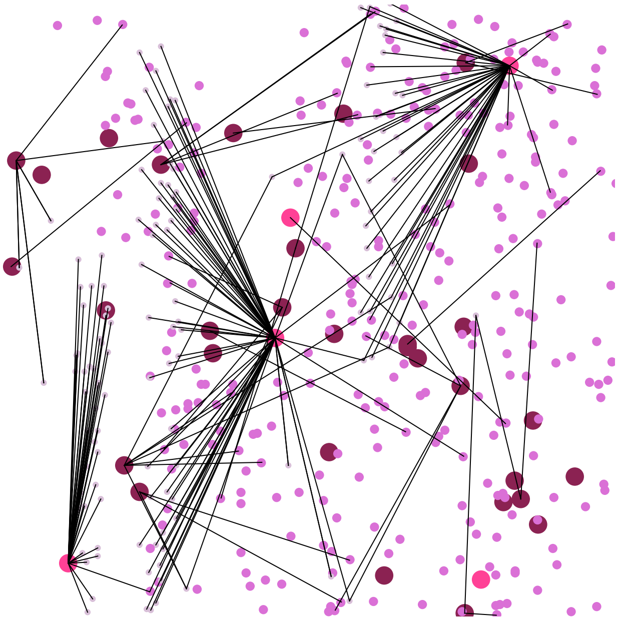

Figure 6: Screenshot of RetSim during a step

In RetSim we do not make any distinction between customers that are part of the membership programme or not. RetSim assumes that all the customers are members. This give us a way to track individual behaviours of all customers, which is beneficial.

The output of RetSim is a CSV file that contains the fields: *Step*, *Type* of *Transaction* (e.g. one sale, three returns), *Customer Id*, *Sales Clerk Id*, *Sales Price*, *Item Id* and *Item Description*.

7.1. Scenarios simulated

We aimed to perform a simulation that would produce a comparable data set to our sample data set which contained 36 sales clerks and around 45000 receipts and 81500 articles sold. The simulation was loaded with a subset of about 11000 articles from the real store.

We ran RetSim for 361 steps (working days of the store), several times and calibrated the parameters given in order to obtain a distribution that get closer enough to be reliable for testing. We collected several log files and selected three from the latest runs. Table 5 compares three runs of RetSim against the original data. Since this is a randomised simulation the values are of course not identical.

Table 5: Statistical Analysis Store one vs RetSim Simulations

Statistic	Store 1	RetSim1	RetSim2	RetSim3
Articles sold	81441	103716	95847	96492
Avg. Sales Price	372.3	405.5	405.2	407.1
Std. Sales Price	510.9	555.1	550.7	552.2

7.2. Social Network Calibration

We experimented with calibrating our results and aim to simulate the network presented in section 4.2.. Our aim was to obtain approximately the same amount of nodes and edges. We used the out-degree distribution to associate sales clerks with customers. So each sales clerk is capable to handle more or less customers during each step of the simulation and this creates the difference between nodes. This difference is interpreted in the real world by two parameters. The first is how many days a sales clerk work and the second is how good sales clerks they are. Accordingly, we only allow sales clerks with a high *in-degree* to be active during most of the steps. It means that we deactivate some sales clerks during any one specific step.

After several experimental runs and around 180 steps, keeping the most of the parameters from the original simulation, we selected one of the simulation runs to show in table 6.

Table 6: Network Simulated

Statistic	RetSim
Nodes	4948
Edges	5339
Sales Clerks	36
Customers	5303
Avg. Degree	1.079
Avg. Weighted Degree	499.1
Modularity Undirected	0.845
Diameter Undirected	8
Avg. Path Undirected	4.19

7.3. Evaluation of the model

We start the evaluation of our model with the verification and validation of the generated simulation data (Ormerod and Rosewell, 2009). The verification ensures that the simulation correspond to the described model presented by the chosen scenarios. We described RetSim in section 5.. In our model, we have included several characteristics from a real store, and successfully generated a distribution of sales that involved the interaction of sales clerks and customers. However, there are a few characteristics left from the real model such as discounts.

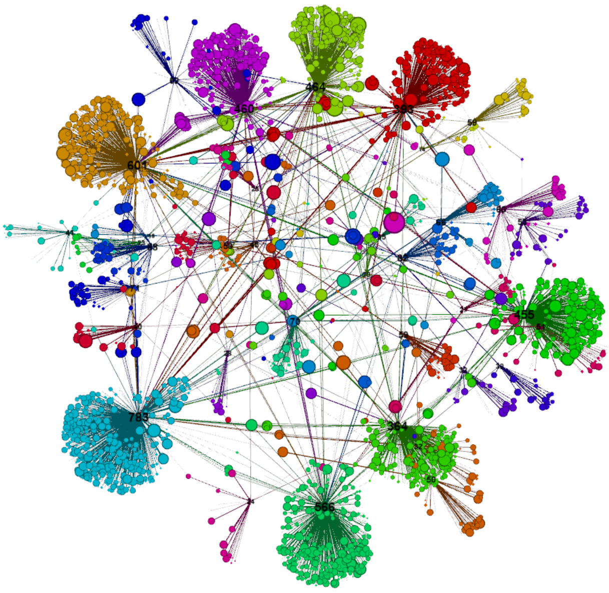

Figure 7: Small Simulated network

The validation of the model answer the question: *Is the model a realistic model of the real problem we are addressing?* After several runs of the simulation to calibrate it, we are able to answer that question affirmatively. We present some generated distributions of sales that are comparable visually in figure 8, 9 and 10.

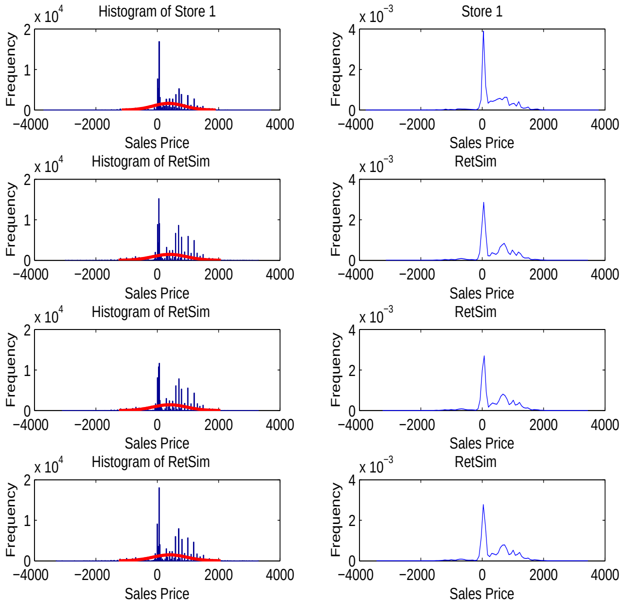

Figure 8: Comparison of simulated vs real data

Figure 8 shows a comparison of RetSim and the real sample data extracted from store one. We note several things: first the shape of the distributions look similar. Before zero are all the returns with a shape of a flat normal distribution. Between zero and 100 are the most frequently sold items such as shoe laces or accessories, which produces a peak. After 100 and before 2000 is the most common rank for shoes, so it presents another part of the distribution that contains the mean.

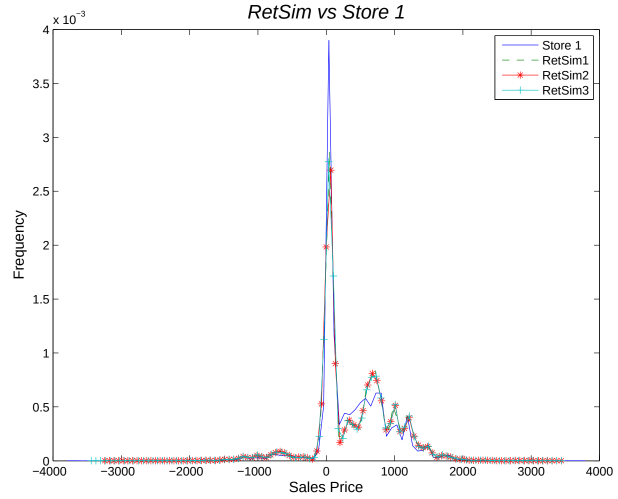

Figure 9: Comparison of distribution of simulated vs real data

Figure 9 shows an overlap of our sample store with different simulation runs by RetSim. Visually the distributions look similar. However there are several differences in the small shapes.

In figure 10 we can see a box plot comparison of store one with the RetSim runs. We can visually identify that the five statistical measures provided by the box plot are similar without being identical.

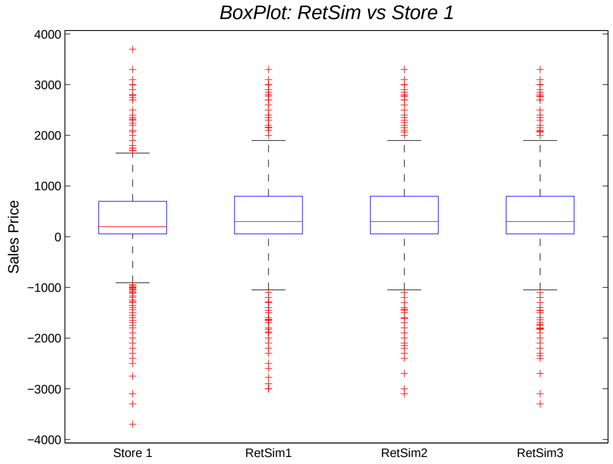

Figure 10: Box plot of simulated vs real data

Now we will focus on evaluating the simulated network presented in section 7.2.. The simulation in comparison with the original data seems visually very similar. There are similarities between the hub topology, number of nodes, and sales clerks. However we also find some dissimilarities between the weighted average degree, which in the simulation was below the original data.

There is more homogeneity between the purchases of the customers in the original data than in the simulated data. This could be due to the random nature of the selection of items in the simulation. Notice the visual differences between figure 3 and 7.

Another difference that we found is that the simulated network generates one single giant component. In the original data we could perceive a few sales clerks that perhaps just worked there for a single/few days and only served few customers. Those sales clerks are identified as islands and separated components. The analysis of these islands might be of interest for fraud detection.

We can also look at the modularity of the simulated network as an emerging behaviour of the customers. Both, the original and the simulated network are very similar and build their communities around the sales clerks. This can be clearly visualized by the different colours used in all the visualizations.

So in summary, our agent model with its programmed micro behaviour, produces the same type of overall interaction network that we can observe in the original data, and furthermore, this interaction network give rise to the same macro behaviour for the whole store as for the real store as well.

Since we are running a simulation we argue that the differences are not significant for our purpose, which is to use this distribution to simulate the normal behaviour of a store, and later combine this with injected anomalies and known patterns of fraud.

8. CONCLUSIONS

RetSim is a simulation of a retail shoe store with the objective to generate a sales data set that can be used for research into fraud detection. Synthetic data sets generated with RetSim can aid academia, companies and governmental agencies to test their methods or to compare the performance of different methods under similar conditions on the same test data set.

In section 3. we formulated our research question for this paper: *How could we model and simulate a retail shoe store and obtaining a realistic synthetic data set for the purpose of fraud detection?* In section 5. we presented the RetSim model, which is based on the ODD methodology. In order to better support our claim and answer our research question we analysed the type of data needed to generate and output as a CVS file (see section 7.) and we evaluated and verified our model in section 7.3..

It is important to know how much information from the real data set is contained in the generated synthetic data. First we do not keep any record of who is purchasing anything in the store, we based our simulation purely on statistical measures and network measures that give us an approximate description of how the individual agents behave. This means that the retail store can be sure that the privacy from the customers is preserved when using RetSim.

We argue that RetSim is ready to be used as a generator of synthetic data sets of commercial activity of a retail store. Data sets generated by RetSim can be used to implement fraud detection scenarios and malicious behaviour scenarios such as a sales clerk returning stolen shoes or unusually low productivity of a sales clerk during a specific day which could mean that the clerk is not entering some of the receipts into the system. We will make a stable released of RetSim available to the research community together with standard data sets developed for this article and further research.

For future work we plan several improvements of and additions to the current model. RetSim can be calibrated to improve the results presented in section 7. and make the data set more realistic.

In order to generate records with malicious behaviour we plan to extend RetSim to also generate malicious activity that can come from the sales clerk, customer or even the managers, or combinations of these.

Among the additions we consider are: inventory control, discounts and promotions that affect the demand of certain products. We can also add hidden parameters to sales clerks such as skills in sales, which will increase the number of customers and the average cost of items purchased. Another possible inclusion in future versions is an interesting behaviour, the self transaction, where a sales clerk can play the role of a customer and a sales clerk at the same time. This behaviour can play a key role in order to find cases of fraud.

ACKNOWLEDGMENTS

The first author wants to thank his wife Helena Lopez for bringing life to his daughter Linnea just a couple of weeks after this work was submitted. Both wife and daughter helped as inspiration and motivation for this work.

REFERENCES

Naoki Abe, Bianca Zadrozny, and John Langford. Outlier detection by active learning. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06*, page 504, 2006. doi: 10.1145/1150402.1150459.

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

Mathieu Bastian, Sebastien Heymann, and M Jacomy. Gephi: An open source software for exploring and manipulating networks. *International AAAI conference on ...*, 2009.

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 - Multi-agent simulation system for monitoring retail activities. pages 20--26, 2008. ISBN 8890073268;978-889007326-7.

Volker Grimm, Uta Berger, Finn Bastiansen, Sigrunn Eliassen, Vincent Ginot, Jarl Giske, John Goss-Custard, Tamara Grand, Simone K. Heinz, Geir Huse, Andreas Huth, Jane U. Jepsen, Christian Jø rgensen, Wolf M. Mooij, Birgit Müller, Guy Pe'er, Cyril Piou, Steven F. Railsback, Andrew M. Robbins, Martha M. Robbins, Eva Rossmanith, Nadja Rüger, Espen Strand, Sami Souissi, Richard a. Stillman, Rune Vabø, Ute Visser, and Donald L. DeAngelis. A standard protocol for describing individual-based and agent-based models. *Ecological Modelling*, 198(1-2):115--126, September 2006. ISSN 03043800. doi: 10.1016/j.ecolmodel.2006.04.023.

Yifan Hu. Efficient and High Quality Force-Directed Graph. *The Mathematical Journal*, 10:37--71, 2005.

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 and Stefan Axelsson. Money Laundering Detection using Synthetic Data. In Julien Karlsson, Lars ; Bidot, editor, *The 27th workshop of (SAIS)*, pages 33--40, Örebro, 2012a. Linköping University Electronic Press.

Edgar Alonso Lopez-Rojas and Stefan Axelsson. Multi Agent Based Simulation (MABS) of Financial Transactions for Anti Money Laundering (AML). In Audun Josang and Bengt Carlsson, editors, *Nordic Conference on Secure IT Systems*, pages 25--32, Karlskrona, 2012b.

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

Paul Ormerod and Bridget Rosewell. Validation and Verification of Agent-Based Models in the Social Sciences. In Flaminio Squazzoni, editor, *LNCS*, pages 130--140. Springer Berlin / Heidelberg, 2009. ISBN 978-3-642-01108-5.

Clifton Phua, Vincent Lee, Kate Smith, and Ross Gayler. A comprehensive survey of data mining-based fraud detection research. *Arxiv preprint arXiv:1009.6119*, 2010.

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.

Grant Thornton. Reviving retail Strategies for growth in 2009 Executive summary. Technical report, Grant Thornton, 2009.

AUTHORS BIOGRAPHY

MSc. Edgar A. Lopez-Rojas

Edgar Lopez is a PhD student in Computer Science and his research area is related with 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. In 2011 he obtained a Masters degree in Computer Science from Linköping University in Sweden.

Dr. Stefan Axelsson
Stefan Axelsson is a senior lecturer at Blekinge Institute of Technology. He received his M.Sc in computer science and engineering in 1993, and his Ph.D. in computer science in 2005, both from Chalmers University of Technology, in Gothenburg, Sweden. His research interests revolve around computer security, especially the detection of anomalous behaviour in computer networks, financial transactions and ship/cargo movements to name a few. He is also interested in how to combine the application of machine learning and information visualization to better aid the operator in understanding how the system classifies a certain behaviour as anomalous. Stefan has ten years of industry experience, most of it working with systems security issues at Ericsson.

Dan Gorton, Licentiate of Engineering
Dan Gorton is a Ph.D. candidate at KTH Royal Institute of Technology. He received his M.Sc. in computer science in 1997 at KTH Royal Institute of Technology, in Stockholm, Sweden, and a Licentiate of Engineering in computer engineering in 2003 at Chalmers University of Technology, in Gothenburg, Sweden. His current research focuses on risk management of online financial services, including fraud detection. Previous research has focused on extending intrusion detection with alert correlation and intrusion tolerance. Dan has 15 years of industry experience, working with different security and risk issues primarily within the banking, defense, and telecom sectors.