Using Computer Simulation for Research in Financial Fraud

Edgar Lopez-Rojas, PhD
Post-doctoral researcher at NTNU in Gjovik, Norway

May 10, 2017

NTNU
Norwegian University of
Science and Technology

Introduction

Simulating Financial Transactions

Case studies

Conclusions

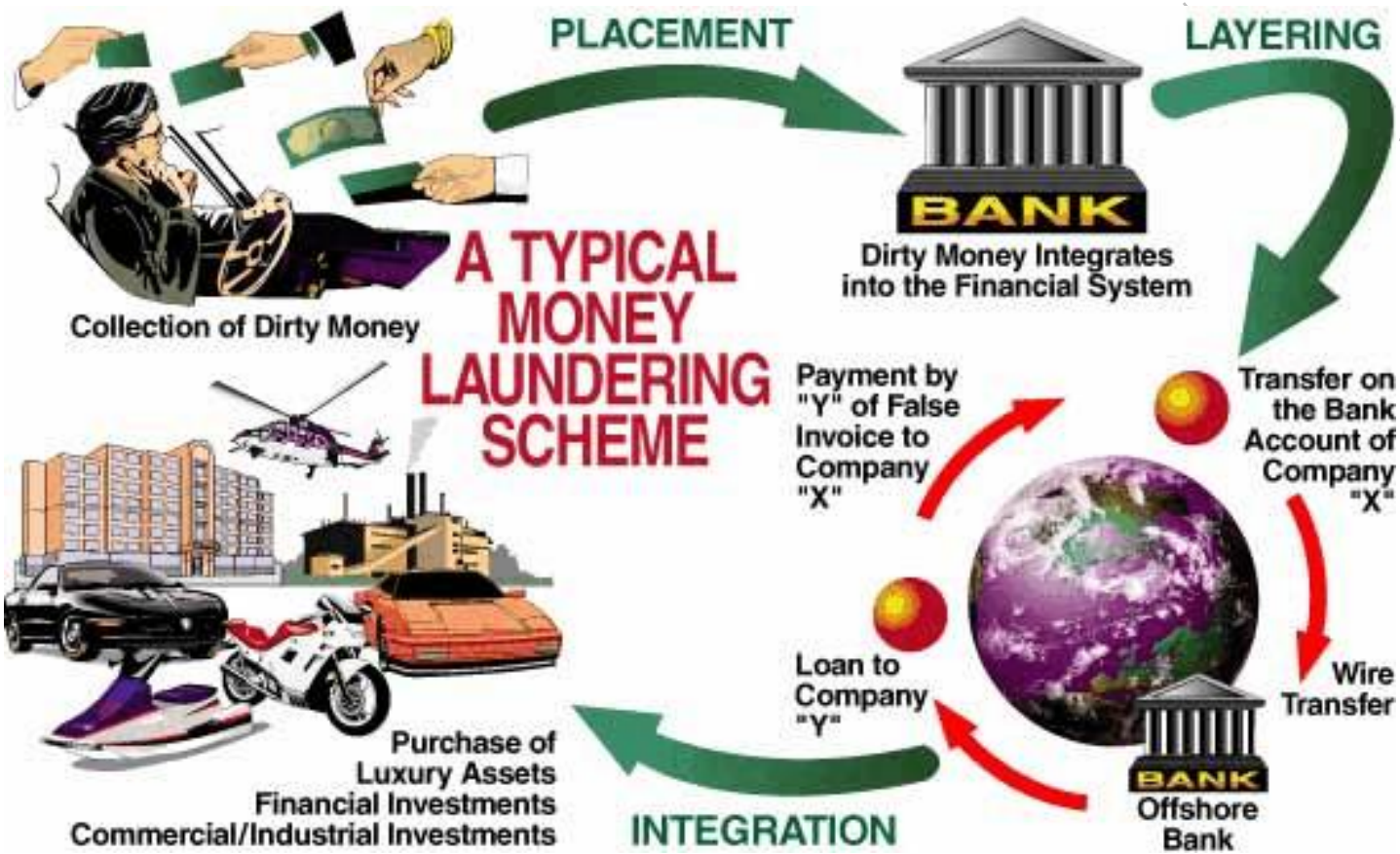
4

Financial Fraud Detection

- Financial fraud can be defined as an intentional act of deception involving financial transactions for personal gain.
- The result of this is a financial loss for a corporation or a person.
- Many financial institutions have implemented controls to prevent fraud.
- But most of these controls are abused by criminals
- When prevention fails, the victim is responsible for reporting the fraud to the financial institution.
- Money Laundering is a particularly complex case of financial fraud and it consists of disguising the proceeds of criminal activities, to make them appear as if they originated from a legitimate source

NTNU
Norwegian University of
Science and Technology

Money Laundering



NTNU
Norwegian University of Science and Technology

The problem of applying effective controls

- By law, financial institutions protect the financial information of their customers; but at the same time they need to control and report suspicious or fraudulent behaviour.
- If you are not working inside a financial institution, then it is difficult to obtain financial datasets for testing these controls.
- Financial Institutions have internal policies to protect customer data, even from their own employees.
- Even inside a financial organization, it is difficult to develop effective controls without going through many cycles of trial and error with business operations.
- Finally, researchers in the specific field of financial fraud encounter many during their research.

NTNU
Norwegian University of
Science and Technology

7

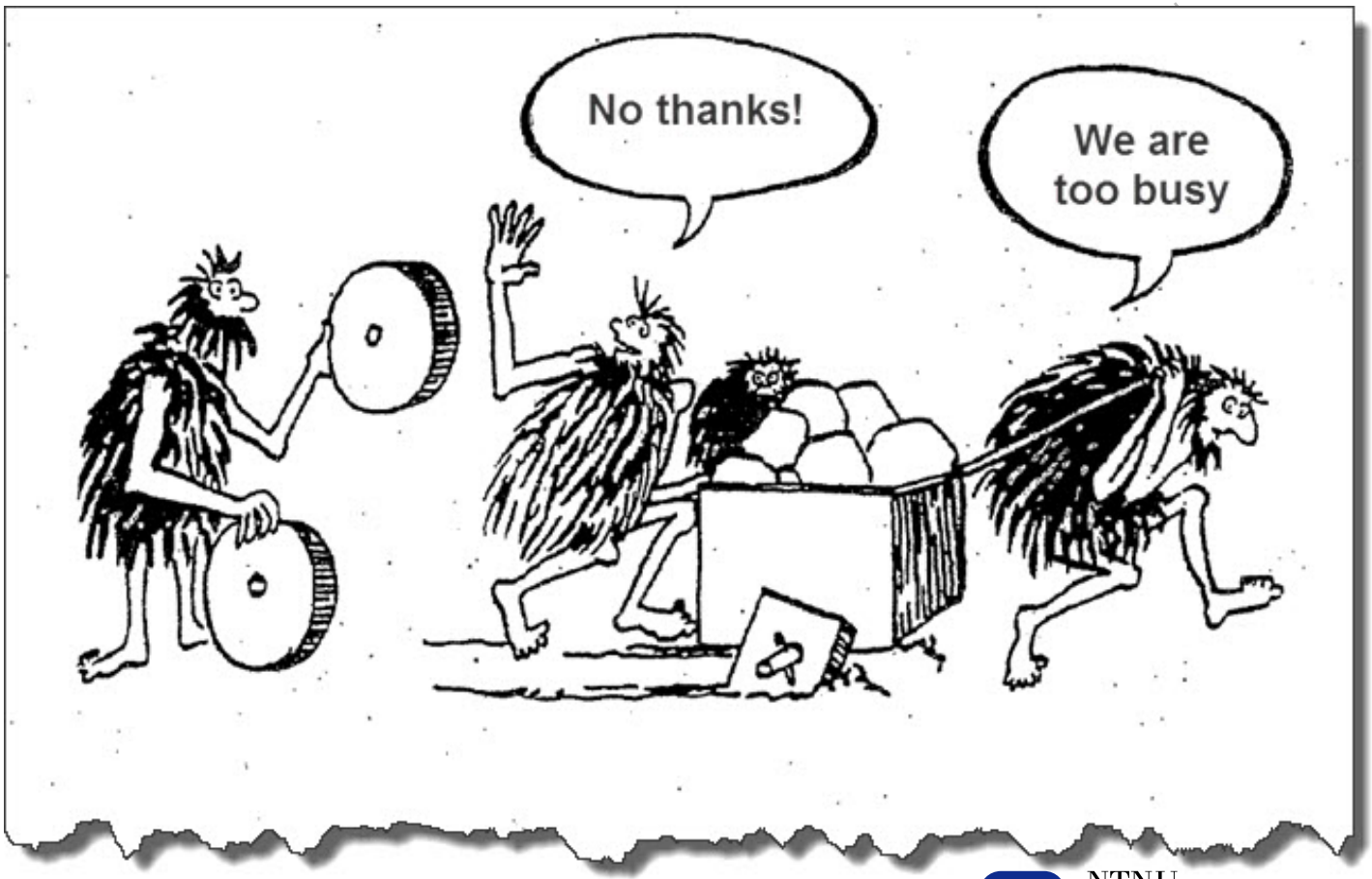
No Data ... What to do now?

- Several years ago, we started to address some of the problems in developing novel methods for fraud detection in general and for fraud detection for Mobile Money Payments in particular.
- Unfortunately, this Mobile Money Service was not producing any data to further our research.
- Instead of waiting for the data to come, we started to develop a data generation method that is based on simulation of financial transactions.
- We worked within the domains of retail sales and financial services.
- We used computer simulations for research in the area of financial services.

NTNU
Norwegian University of
Science and Technology

8

I have a tool for you



NTNU
Norwegian University of Science and Technology

9

Simulator ... Simulation

- Computer Simulation is used in many different fields and domains to infer conclusions about the behaviour of real-world phenomena.
- For example: weather prediction, logistics, thermodynamics, electronics, etc.
- It is a relatively new development for fraud detection in financial services,
- mainly due to the privacy issues explained earlier.
- Researchers had little access to data, while law enforcement authorities had other concerns (preventing crime).

NTNU
Norwegian University of
Science and Technology

Multi-Agent Based Simulation (MABS)

- Multi-Agent Based Simulation (MABS) is a specific type of Computer Simulation
- MABS are built from the bottom up.
- The design doesn't need to know the complex structural behaviour or emerging behaviour that it is simulating.
- It uses knowledge of individual behavior and then creates collectives of these individuals: these collectives exhibit complex emerging behaviour.
- This makes it very useful to model financial services behaviour that emerges from these collections of individual customers; this emergence produces complex financial interactions.
- MABS keywords: Agents, Environment, States, Behaviour, Steps

NTNU
Norwegian University of
Science and Technology

11

Why Synthetic Data?

There are many benefits of using synthetic datasets:

- Data is ready and available
- Privacy of customers is not affected
- Results can be disclosed to, and compared by, other researchers
- Different scenarios can be modeled using well controlled parameters.
- We can avoid the class imbalance problem for ML classification.
- Prior knowledge of the interesting (fraud) cases avoids the problem of mislabeled classes for ML.

NTNU
Norwegian University of
Science and Technology

Why Synthetic Data?

Unfortunately, there are still issues that arise when using synthetic datasets:

- Data generated might not be representative or realistic
- Data can be biased
- It is difficult to build a fully realistic model, due to the complexity of the variables and parameters.
- Any fraud detected can be used only as an example and does not represent a real case of fraud.
- The methods discovered using synthetic datasets still need to be retested with real world datasets.

A Financial Simulator is our Tool

- We realized that simulation was needed for fraud research into financial transactions.
- Like an astronomer, we started to build our own telescope to see the stars.
- We built 3 simulators that worked as our ”telescope” for research in financial transactions: RetSim, BankSim and PaySim.
- We developed a method to take advantage of our tool to produce synthetic datasets and to address most of the issues that arise with it.

NTNU
Norwegian University of
Science and Technology

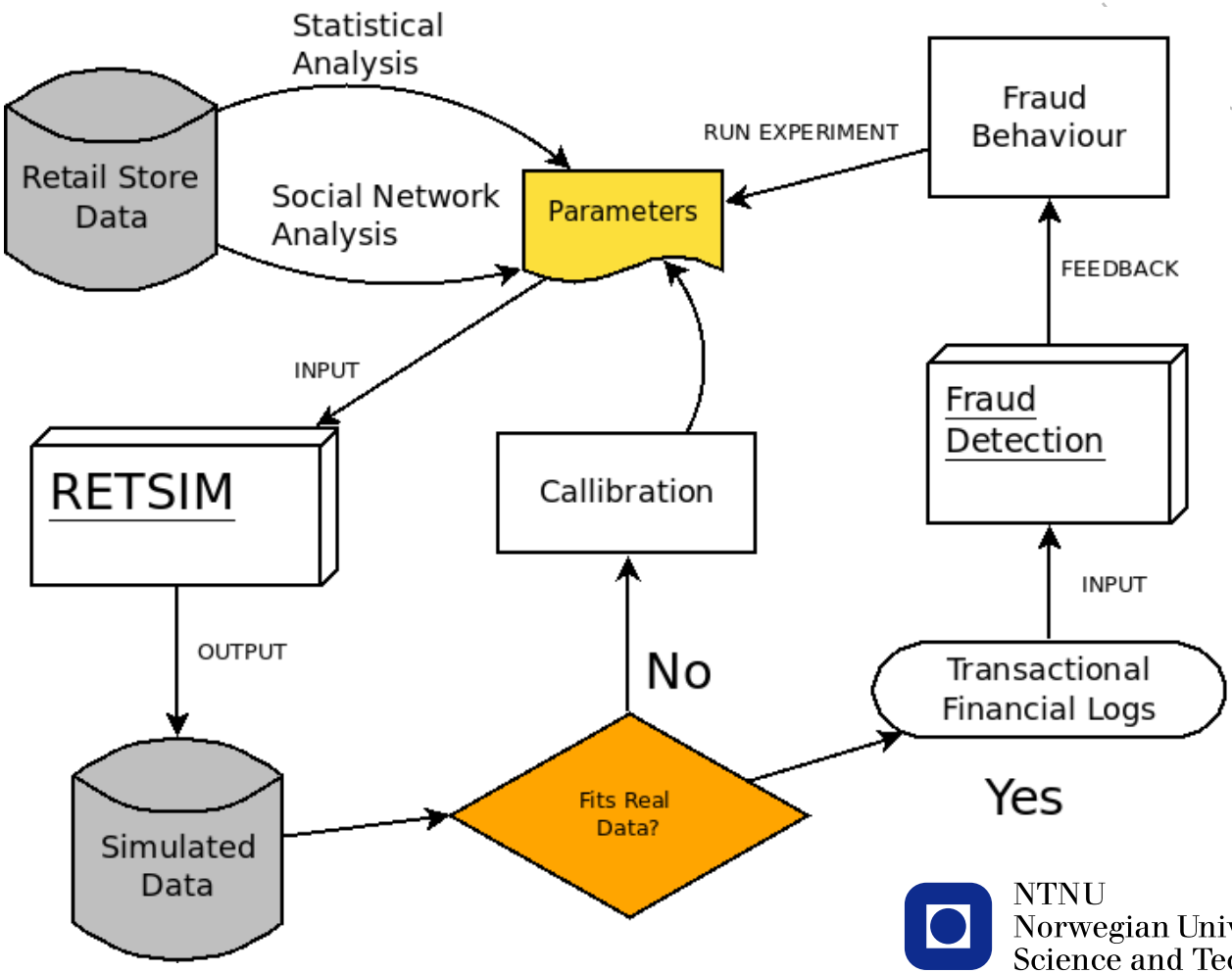
Simulating Financial Transactions

As a summary our method follows these steps in order to simulate financial transactions and perform experiments:

1. Obtain a sample of real data.
2. Perform data analysis to extract aggregated information for input into the simulator.
3. Add parametrization for expected fraud scenarios.
4. Run the simulator several times, using different random number seeds and/or different fraud configurations.
5. Apply the fraud detection methods on the generated synthetic dataset.
6. Summarize the results and performance from the experiments.
7. Repeat from step 3 on, for various fraud scenarios.

16

Case Study: RetSim



NTNU
Norwegian University of Science and Technology

17

1. Obtain a sample of real data

   — This is probably the hardest part of the whole method
   — Real data is important, because without it the simulation results can not be trusted.
   — Fraud detection results are highly dependent on the dataset.
   — For better results, a sample dataset that represents the financial service is required
   — This means that it covers enough (interesting) periods of time to learn more during the data analysis.
   — If there is fraud, it should be properly labeled and identified with respect to which class of fraud it belongs in.

NTNU
Norwegian University of
Science and Technology

18

1. Obtain a sample of real data

The real data sample can be obtained in several different ways:

- Full dump of a database (100% access)
- All of the data over a period of time
- Partial attributes of the data for some period of time
- All data anonymised with respect to customer information
- Anonymisation by adding noise corruption (lowers the data quality)
- Simply aggregating information over a period of time

19

2. Perform a data analysis to extract aggregated information for input to the simulator.

- Depending on the way the real data is provided, we need to perform several operations to convert the data into the format required
- The simulator uses aggregations of information over a period of time, as input.
- The time granularity of the aggregation is specified on the simulation as a STEP
- To accurately mimic the data distribution, we must extract aggregated information from the original data that matches each step in the simulation
- There are also initial values and other input values extracted from the real data

NTNU
Norwegian University of
Science and Technology

2. Perform a data analysis to extract aggregated information for simulator input.

- The information extracted is represented in terms of probabilities to ease the decision processes of the agents.
- Social Network Analysis (SNA) helps to recreate the topology of the customers' relations inside the simulation.
- The agent interacts with other agents within the environment and this interaction is specified by the information extracted during the SNA
- The data analysis can also be done by employees of the financial institution that have access to the sample
- Researchers only need the output of this step, to continue the process, this allows financial institutions to preserve the privacy of the customers

NTNU
Norwegian University of
Science and Technology

3. Add parametrization about expected fraud scenarios

— The simulators are usually built to serve a purpose.
— Our simulators contain agents that, under certain conditions, act contrary to the law.
— The synthetic dataset has the benefit that can be generated according to the researcher's needs to study how certain fraud might affect a specific scenario.
— It can be a representation of the original dataset (sample).
— That is why we extract the aggregated information from the sample.
— Part of the simulator validation is to show that, given certain parameters, we can reproduce similar datasets.

NTNU
Norwegian University of
Science and Technology

4. Run the simulator several times using different random number seeds and/or different fraud configurations.

- In order to perform research in this field we need to be able to test different configurations
- The Financial Simulator can also be used to answer all the ”WHAT IF” questions that are common during research
- Researchers can run the simulator several times, using controlled variation on the parameters, to create new scenarios with normal and fraudulent data.
- This is specifically useful for answering questions such as WHAT IF: There is no fraud, there is little fraud, there is a lot of fraud, double the number of customers, and so on.

NTNU
Norwegian University of
Science and Technology

23

5. Apply the fraud detection methods on the generated synthetic dataset

- This is one of the most important steps in the method.
- By changing the parameters in the previews step, we can generate diverse scenarios
- These scenarios produce datasets with data that are labeled as fraudulent or not fraudulent (as appropriate).
- Once a dataset is generated, different methods for fraud detection can be tested and evaluated using the fraud label.
- A method for fraud detection can also be tested and evaluated with different scenarios that use the same fraud label.
- Fraud prevention methods can be also be added to the simulator to test and evaluate against fraud scenarios with the same flagged fraud.

NTNU
Norwegian University of
Science and Technology

6. Summarizing the results and performance from the experiments

- The biggest advantage of using a simulator over a real dataset is that we know with certainty how much fraud is present and where it is located.
- In a real dataset, it is impossible to guaranty that there isn't any undetected hidden fraud.
- Since we control our malicious agents, we can flag *all* fraudulent behaviour, because we have prior knowledge about the level of fraud injected into the dataset.
- Measuring all the fraud present in a dataset is one of the biggest challenges when using real data, but not with synthetic data.

25

7. Repeat from step 3 for different fraud scenarios
   - After analysing the results on the previous step new questions may arise
   - New scenarios can be generated
   - Fraud detection methods can be modified to improve the results
   - A simulator can be used in a loop to improve results and perform research in fraud detection
   - Re-Starting at step 3 is more effective for research but some researchers might chose to work on a previously generated dataset (step 5) to test different methods and compare results against previous research.

NTNU
Norwegian University of
Science and Technology

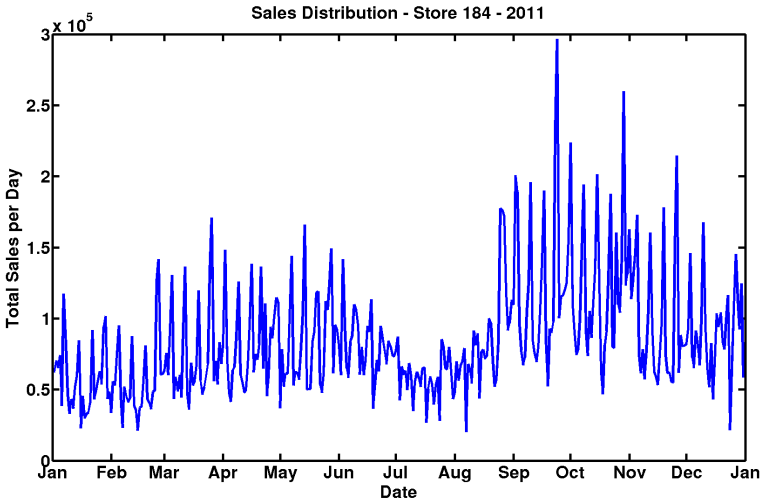
27

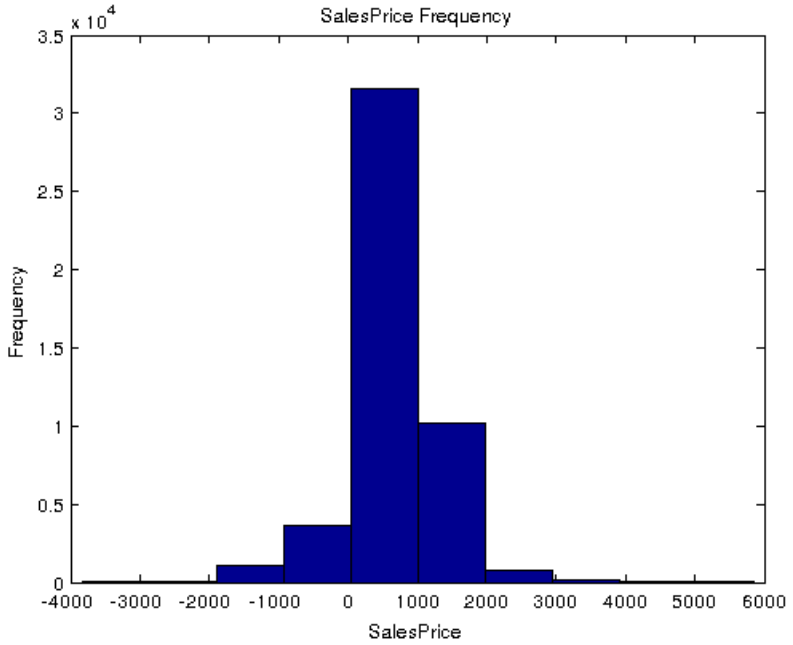
The RetSim Simulator

- RetSim (Retail Store Simulator) is an agent-based simulator of a shoe store
- It is based on transactional data from one of the largest retail shoe sellers in Sweden
- We received access to a full dump of the dataset
- During the data analysis step, we selected a store that contained sufficient information about customers to enable performance of SNA to build a network
- One year of transactions was selected as the period of time for analysis.
- Each step on the simulation represents a day of sales on the store.

28

Data Analysis Sales





NTNU
Norwegian University of
Science and Technology

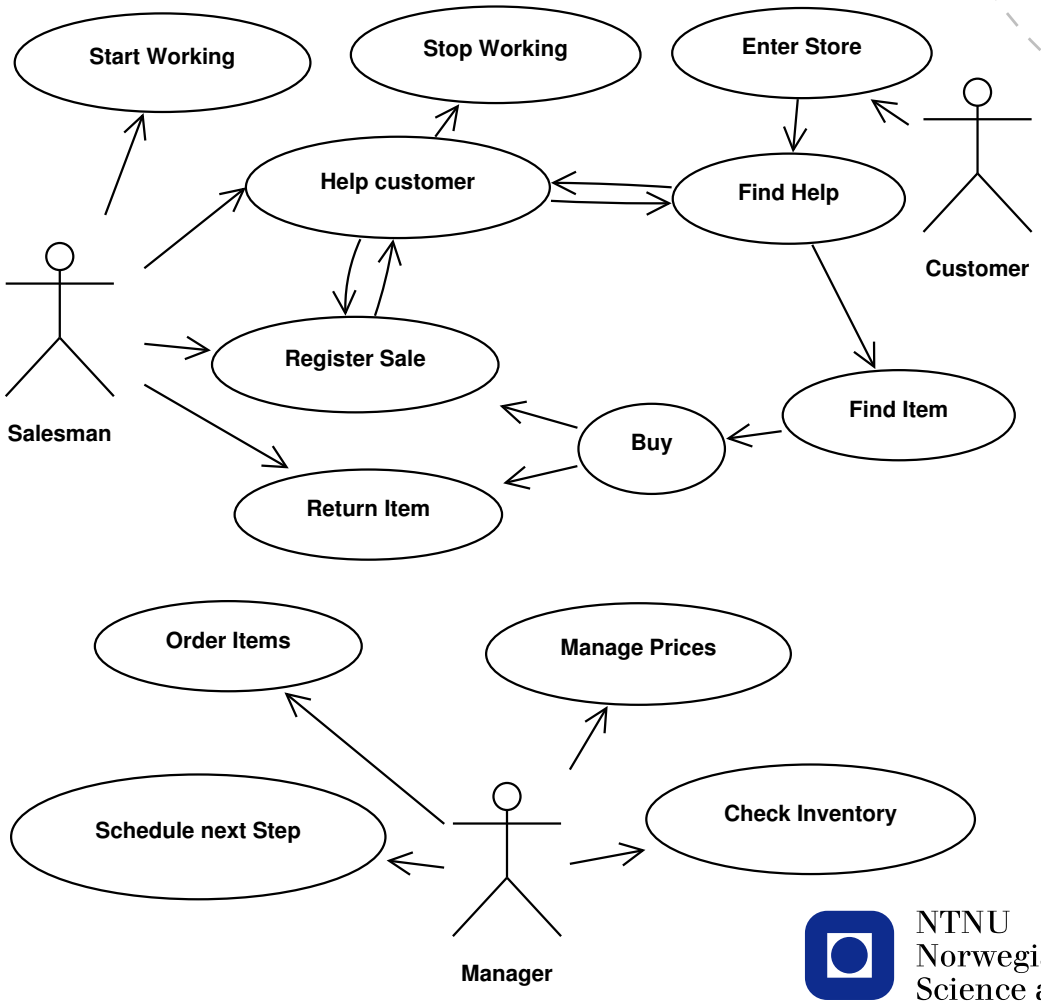
29

Data Analysis SNA

NTNU
Norwegian University of
Science and Technology

30

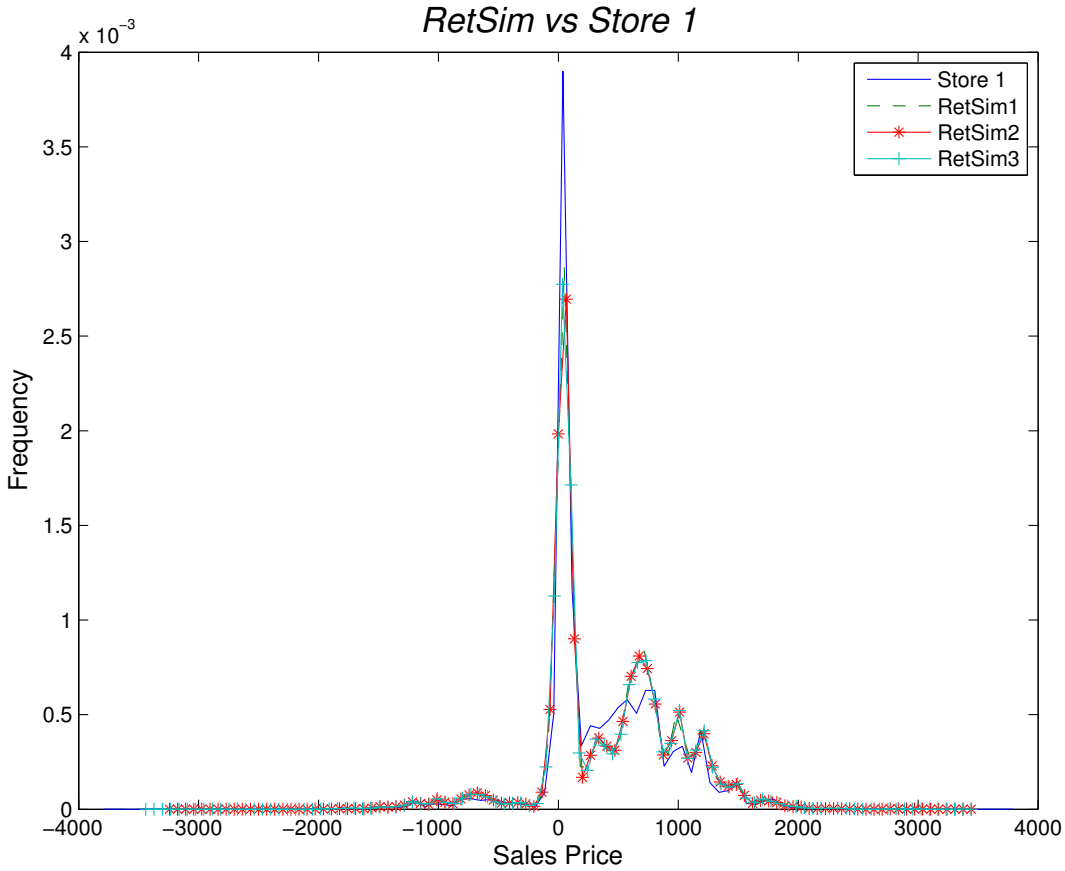
Model



NTNU
Norwegian University of
Science and Technology

31

Calibration



NTNU
Norwegian University of
Science and Technology

32

Fraud Scenarios

We injected 2 fraud scenarios:

- The Refunds scenario includes cases where the salesman creates fraudulent refund slips, keeping the cash refund for him- or herself.
- The refund scenario was simulated by estimating the average number of refunds per sale and the corresponding standard deviation
- Coupon reductions/discounts scenario 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.

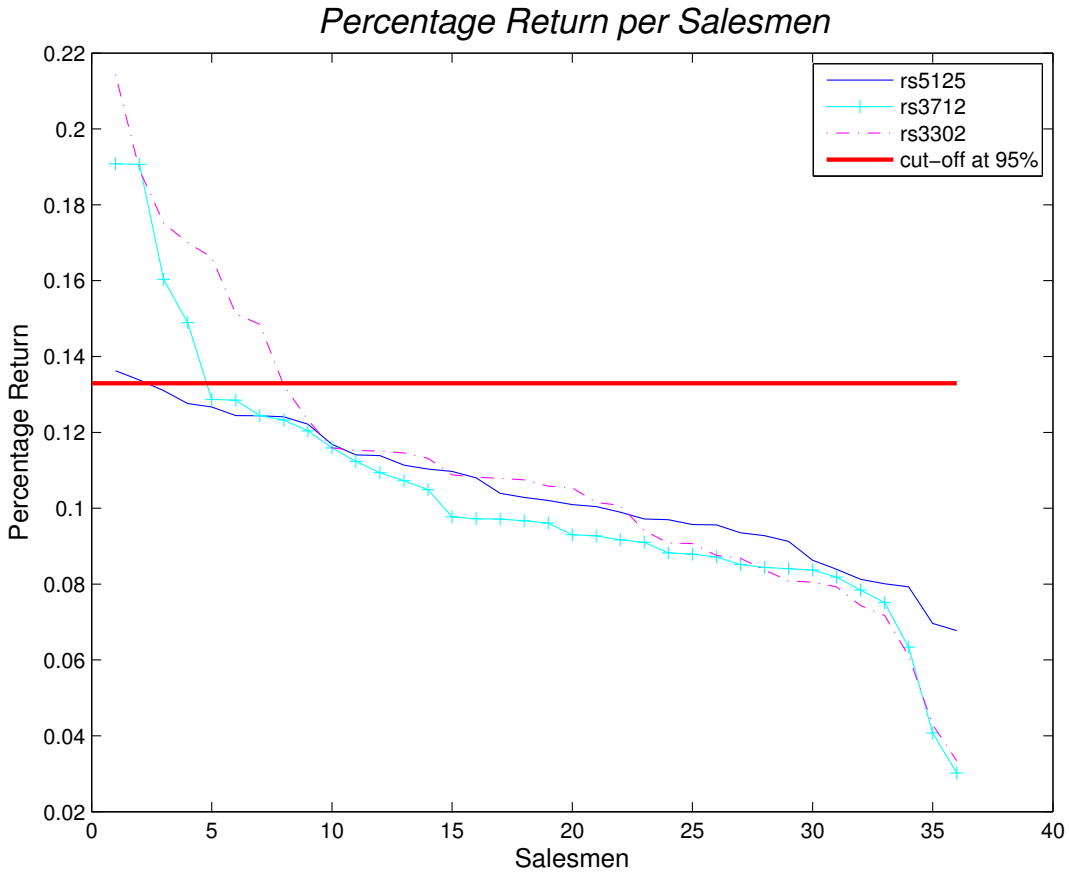
33

Thresholds everywhere

- We have our ”telescope”, now its time for ”astronomy”.
- One of the more common techniques to detect fraud is the use of limits or thresholds.
- So we used the RetSim simulator to dig into this topic in the fraud scenarios modeled.
- We live in a world where machine learning is assisting in many fields of human endeavor.
- Fraud is one of the areas that ML is being used.
- How effective are threshold methods and what is the margin of improvement from machine learning methods?
- Do ML methods have improved cost over threshold methods?

NTNU
Norwegian University of
Science and Technology

Using Thresholds



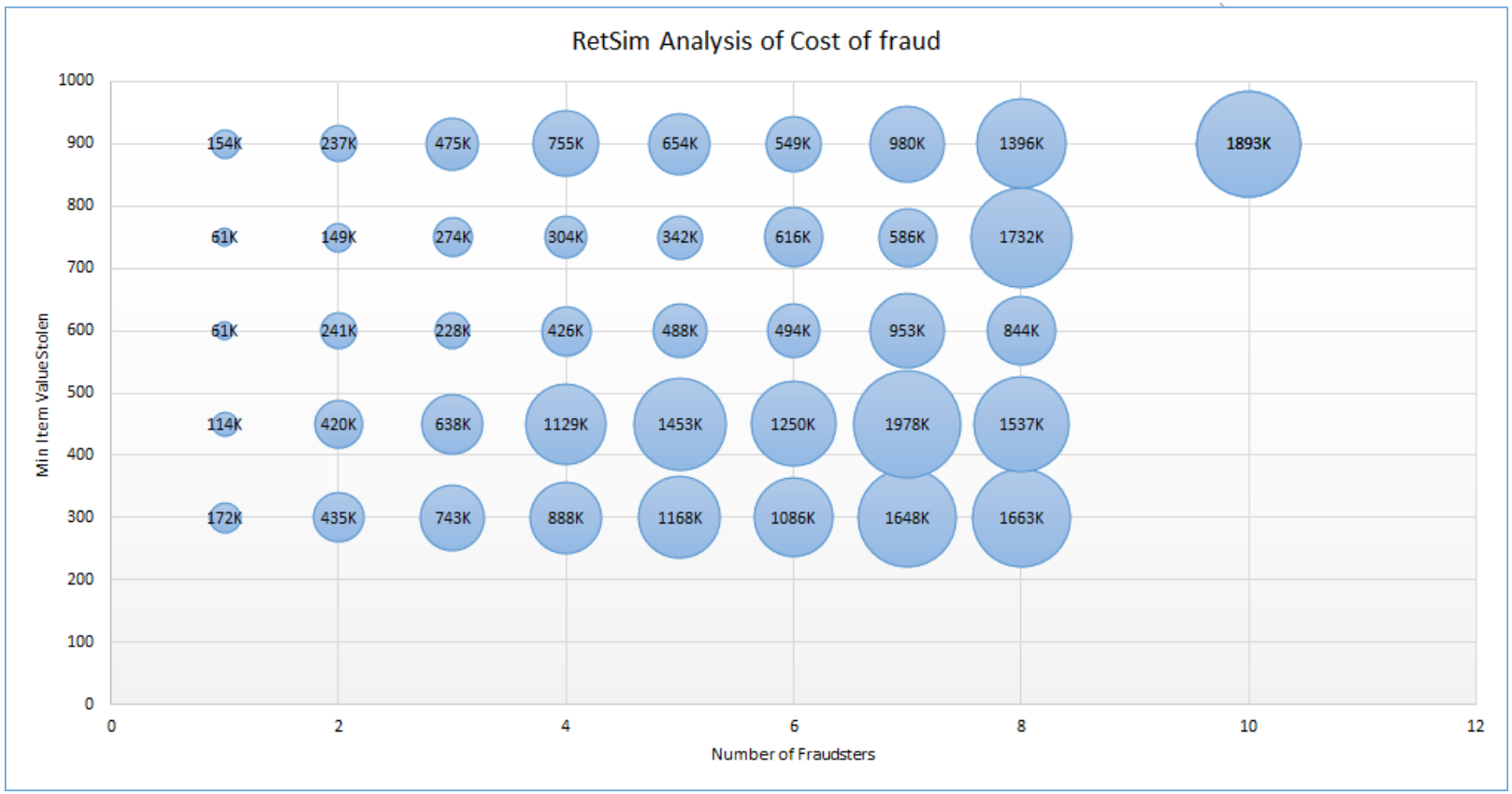
NTNU
Norwegian University of
Science and Technology

RetSim gets more useful ideas

- Why not use it for measuring the cost of fraud?
- Fraud cost is usually estimated by using statistics.
- But it's not very accurate and it cannot predict future scenarios
- We use our "telescope" to explore unseen areas of the "universe".
- This idea of "the cost of fraud" will motivate managers to invest in security

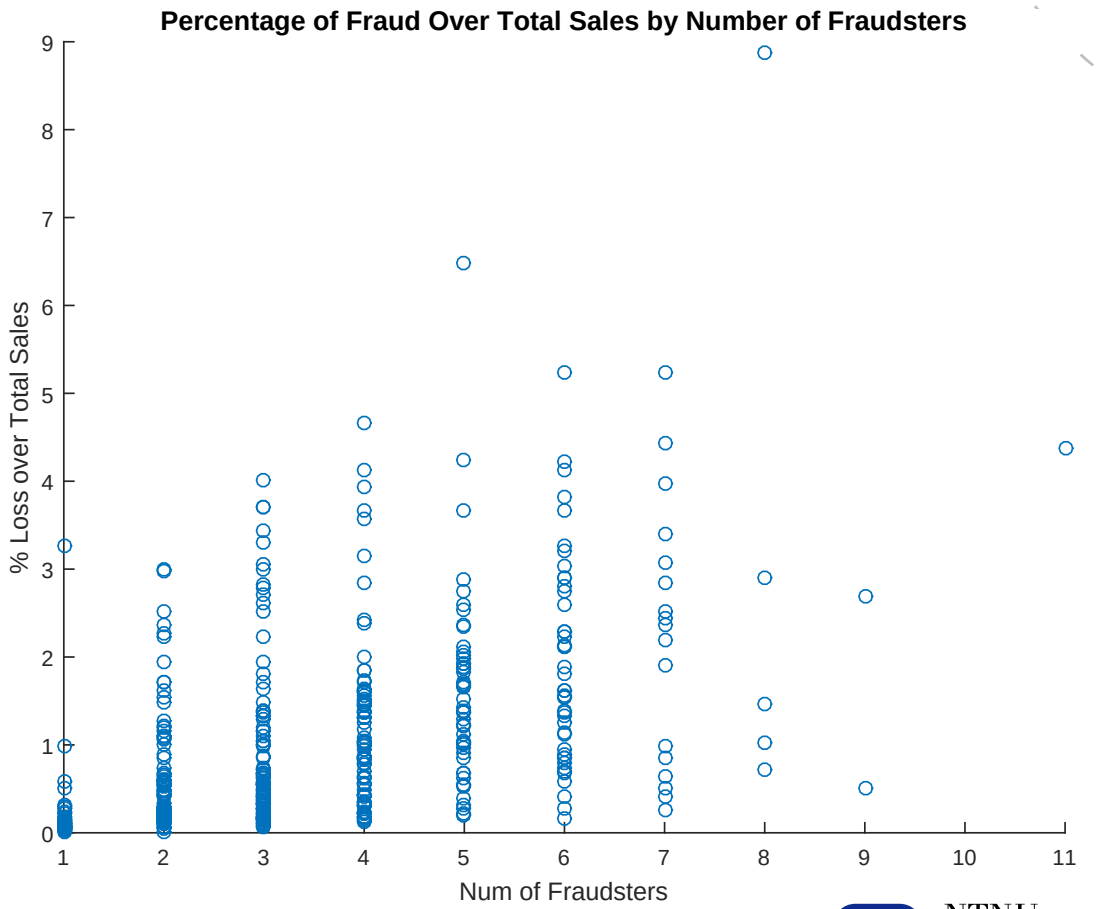
36

Using RetSim to measure the cost of fraud



NTNU
Norwegian University of
Science and Technology

Using RetSim to measure the cost of fraud



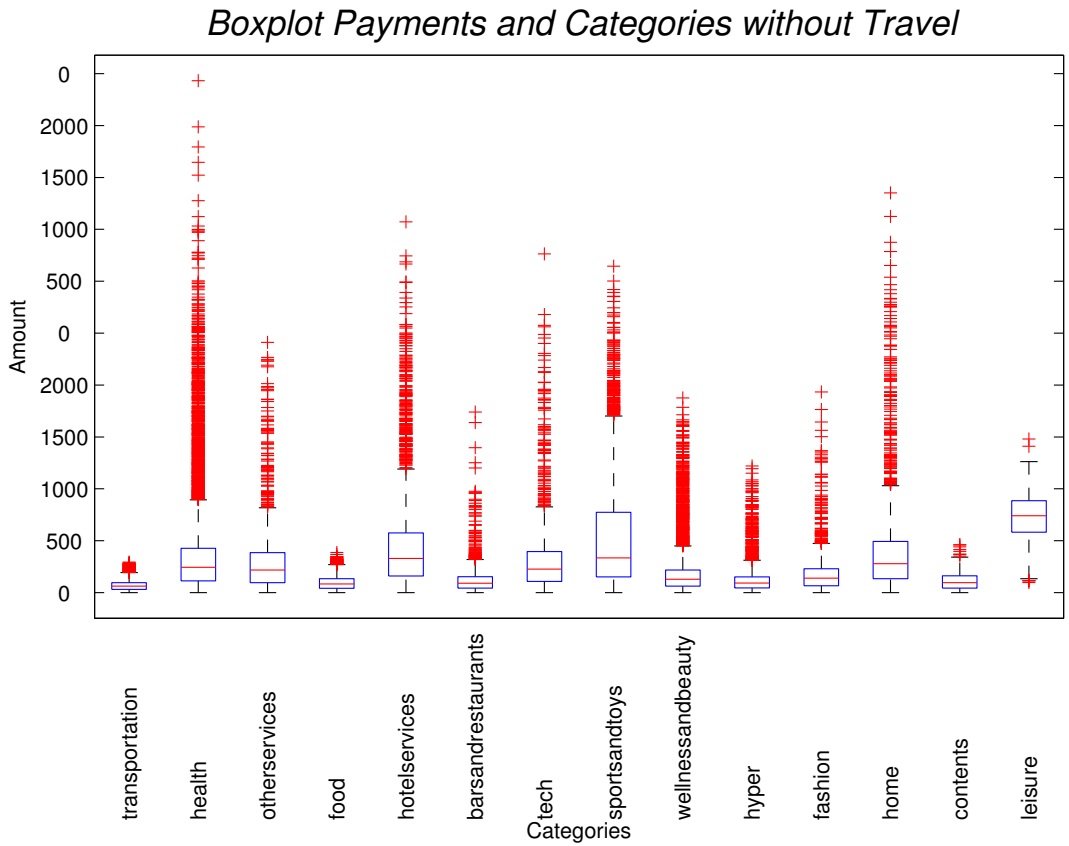
NTNU
Norwegian University of
Science and Technology

The BankSim Simulator

- BankSim is a Bank Payment Simulation for Fraud Detection Research
- BBVA bank in Spain was sharing a webservice to query aggregated financial transactions with the purpose of developing apps for a contest.
- We participated in this contest
- But the first thing we did, was to query all available data and store it locally for our research
- Then we built ”The BankSim Simulator” that used the aggregated information
- The bank never disclosed any information about their customers when it would have been useful for us to build a consumption model.

NTNU
Norwegian University of
Science and Technology

Banksim simulated categories



NTNU
Norwegian University of
Science and Technology

PaySim: Financial Simulator of Mobile Money

- PaySim is a Financial Simulator of Mobile Money for Fraud Detection Research
- This was our most recent simulator; it was built just last year.
- We obtained a data sample with more than 24 million transactions from over a period of 1 month.
- We covered 5 of the most important transaction types: CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER.
- We followed the distributions and simulated the normal behaviour
- We later decided to study the case of fraud that occurs when a customer loses control over their account

NTNU
Norwegian University of
Science and Technology

41

PaySim: Financial Simulator of Mobile Money

- For the criminal to empty the account, he needs to use a merchant to cash out the account
- If there are controls to avoid daily withdraws higher than a threshold, the criminal needs to use several mule accounts.
- We simulated 4 scenarios with different thresholds in order to investigate the effectiveness of a control that prevents the emptying of an account.
- If the accounts executes 3 withdraws for the max daily amount, then the account is locked and the customer must contact customer services to unlock the account.

NTNU
Norwegian University of
Science and Technology

PaySim Results

Table 2 Fraud Detection Classification

LogName	Class	Count	Amount	% count	% amount
PS89745 (300k)	FN	27,412	6,724M	1.005%	0.363%
	FP	982	214M	0.036%	0.012%
	TN	2,607,642	1,816,764M	95.579%	98.162%
	TP	92,211	27,076M	3.380%	1.463%
PS80775 (600k)	FN	24,400	11,291M	0.990%	0.581%
	FP	58	17M	0.002%	0.001%
	TN	2,396,684	1,907,409M	97.239%	98.126%
	TP	43,604	25,114M	1.769%	1.292%
PS00273 (900k)	FN	21,072	12,854M	1.024%	0.768%
	FP	8	1M	0.000%	0.000%
	TN	2,011,006	1,639,699M	97.712%	97.903%
	TP	26,006	22,264M	1.264%	1.329%
PS98516 (1200k)	FN	20,493	16,189M	0.921%	0.858%
	FP	1	0.168M	0.000%	0.000%
	TN	2,186,516	1,849,707M	98.215%	97.993%
	TP	19,248	21,686M	0.865%	1.149%

NTNU
Norwegian University of
Science and Technology

PaySim Results

- If the thresholds are set too high, then the most of the accounts won't be locked during fraud.
- If the thresholds are set too low, then several true customers will see their account locked after normal transactions.

Table 3 Fraud Detection Results

LogName	Precision	Recall
PS89745	98.946%	77.085%
PS80775	99.867%	64.120%
PS00273	99.969%	55.240%
PS98516	99.995%	48.434%

NTNU
Norwegian University of
Science and Technology

45

Sharing our data at Kaggle.com

NTNU
Norwegian University of
Science and Technology

Conclusions

- Fraud detection in financial transactions is affected by the availability of datasets for testing methods
- Our approach presents an alternative: working with synthetic datasets to allow researchers to generate data from diverse scenarios and model fraud.
- Most of our results have been presented with RetSim
- PaySim is just starting to grow and collaboration with other researchers has already been started
- We aim to increase the quality of the synthetic dataset by incorporating more detailed SNA into the parameters
- We aim to use different real datasets to enrich the availability of synthetic datasets

47

This is IT!!!

— Any questions?

NTNU
Norwegian University of Science and Technology