QATAR UNIVERSITY

COLLEGE OF ENGINEERING

DETECTING MARKET MANIPULATION IN STOCK MARKET DATA

BY

HAYA A. AL-THANI

A Thesis Submitted to the Faculty of

the College of Engineering

in Partial Fulfillment

of the Requirements

for the Degree of

Masters of Science in Computing

June 2017

COMMITTEE PAGE

The members of the Committee approve the Thesis of Haya A. Al-Thani defended on 22nd of May, 2017.

_____

Dr. Sumaya A. Al-Maadeed
Thesis/Dissertation Supervisor

_____

Dr. Noora Fetais
Co-supervisor

_____

Dr. Ali Jaoua
Co-supervisor

Approved:

_____

Khalifa Al-Khalifa, Dean, College of Engineering

ABSTRACT

AL-THANI, HAYA, A., Masters : June : 2017, Masters of Science in Computing

Title: Detecting Market Manipulation in Stock Market Data

Supervisor of Thesis: Dr. Sumaya A. Al-Maadeed.

Anomaly Detection is an extensively researched problem that has diverse applications in many domains. Anomaly detection is the process of finding data points or patterns that do not conform to expected behavior within a dataset. Solutions to this problem have used techniques from disciplines such as statistics, machine learning, data mining, spectral theory and information theory. In the case of stock market data, the input is a non-linear complex time series that render statistical methods ineffective. The aim of this thesis, is to detect anomalies within the Standard and Poor and Qatar Stock Exchange using the behavior of similar time series. Many works on stock market manipulation focus on supervised learning techniques, which require labeled datasets. The labeling process requires substantial efforts. Anomalous behavior is also dynamic in nature. For those reasons, the development of an unsupervised market manipulation detection technique would be very interesting. The Contextual Anomaly Detector (CAD) is an unsupervised method that finds anomalies by looking at similarly behaving time series and uses them to predict expected values. When the predicted value is different from the actual value in the time series by a certain threshold, it is considered an anomaly. This thesis will look at the Contextual Anomaly Detector (CAD) and implement a different preprocessing step to improve recall and precision.

ACKNOWLEDGMENTS

This thesis would not have been what it is if not for the support of my supervisors. Thank you, Dr. Sumaya Al-Maadeed, Dr. Noora Fetais and Dr. Ali Jaoua for your advice and motivation. I'd like to also thank Dr. Ali's team who offered their experience and time, Eman Rezk and Fahad Islam.

I would also like to acknowledge the financial advice and background from, Nasser Al-Thani, financial analyst at Qatar Investment Authority.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Anomaly detection is a vast problem that has been researched extensively due to its various application domains. Anomaly detection is the process of identifying data points or patterns in datasets that do not follow expected behavior. These points or patterns can be called anomalies or outliers. This problem has a wide range of applicable domains such as credit cards fraud detection, intrusion detection, healthcare or cyber-security. Anomaly detection is important because anomalies often reflect significant information in a lot of domains. For example, an anomaly found in an MRI image might indicate the presence of a tumor. Anomalies in a network's traffic pattern can mean the computer has been hacked. These many useful applications have made anomaly detection an extensively studied field in statistics and computing. In this thesis, anomaly detection will be used to identify market manipulations in stock market data.

1.1 Anomaly Detection Background

A simple approach to anomaly detection would be defining a range in the data that specify expected normal behavior. Any data point that is outside the normal region is therefore an anomaly. However, several factors make this simplistic approach very challenging:

- Defining a range that includes every possible instance of normal behavior is a difficult, if not impossible task. Often times, what distinguishes normal and abnormal behavior is not so precise. This would make abnormal behavior close to

the boundary very hard to identify.

- Modern malicious actions often evolve to replicate normal behavior. Anomalies resulting from these malicious actions are hard to identify.
- Normal behavior also keeps changing. What is defined as normal behavior now might not be representative in the future.
- Anomalies are defined differently for different applications. For example, in healthcare, any small deviation is significate, such as fluctuations in body temperature. While on the other hand, similar deviations in the stock market might be taken as normal. This makes applying one technique to multiple domains difficult.
- The unavailability of labeled training data for anomaly detection is often a challenge.
- Noise in the data is usually similar to actual anomalies. Distinguishing between the two is a challenge.

To address the above challenges researches explored solutions from a variety of disciplines. These concepts are taken from statistics, machine learning, data mining, spectral theories and information theories. The major factor affecting the selection of anomaly detection technique is the input data itself. The nature of the input data and how instances relate to each other determine the course of action to be taken. Data instances can be sequential data, spatial data or graphical data. In sequential data, the instances are ordered. Examples of these types of data are time-series data, genome data or protein sequences. Data instances in spatial data are related to their neighboring data, for example,

traffic data, geographical data or ecological data. When data is represented as vertices, they are considered graphical data. It is essential to grasp the nature of the data before defining what an anomaly is and how to identify it.

1.2 Stock Market Manipulation Background

This thesis focuses on identifying possible fraudulent activities in stock market data. The stock market is a place where buyers and sellers can trade stocks or securities of publicly listed companies. A stock is a share of a company that represents ownership of a business. A security is proof of ownership of a stock, a bond or any other financial asset. Large companies trade their stocks through an exchange which brings buyers and sellers together in an organized manner. These exchanges exist in many major cities such as the NASDAQ stock exchange and the London stock exchange.

As of 2015, there are 60 stock exchanges world wide with a total market capital of 69 trillion dollars.[1] Regulating the fairness of these markets is a very challenging task. In 2016 alone, the US Securities and Exchange Commission suspended securities trading of 199 issuers to combat market manipulation and fraud threats to investors.[2] Securities fraud is defined as deceptive practices in the trading of stocks or securities. Securities fraud is divided into: broker embezzlement, high yield investment fraud, late-day trading and market manipulation.

---

[1] http://money.visualcapitalist.com/all-of-the-worlds-stock-exchanges-by-size/
[2] https://www.sec.gov/news/pressrelease/2016-212.html

Market manipulation is a big concern for investors. Market manipulation is when the price of a security is artificially inflated or deflated by a group or individual in order to deceive investors and gain profit. A stock price can be manipulated by leaking misleading or incorrect information about a company, limiting the number of shares available to the public or changing trades, quotes or prices in order to create a false demand for a security. Monitoring these malicious actions is very important. A market should ensure integrity, transparency and stability for the benefit of all participants. That is why managing the risks involved with the stock market is such a significant task. A market that is free from all manipulation is crucial for nurturing a stable environment, not only for the benefit of the individuals and companies involved, but for the economic growth of the country itself.

More formally, market manipulation can be defined as any interference with genuine supply and demand of a stock for an illegitimate or deceptive purpose. Market manipulation can come in three different types: information-based, trade-based or action-based manipulation.

Information-based manipulation happens when an individual or group persuades others that the price of a stock is not the actual or current price. The stock can be misleadingly presented as more valuable than what it actually is using false information. An example of this type of manipulation is "pump and dump" manipulation. "Pumping up" is when a seller talks up the value of a stock they are holding. Then the stock is "dumped" or sold into the market at an artificially high price. Another example is when a manipulator "shorts" the stock by spreading false information about the stock to buy it back at a lower price.

Trade-based manipulation is where a manipulator actually completes a transaction with the goal of influencing the stock price. This can be used to make the stock price more favorable for the manipulator. An example of this manipulation is when an oil trader buys all of the supplies of a certain type of oil and keeps them in order to push the price up. Another example is when a buyer and seller agrees to trade at an artificially high or low price. This will influence the price at which other trades occur. A "wash trade" is when a manipulator buys and sells the same stock to mislead other traders in thinking that the stock is more actively traded than what it really is.

Action-based manipulation is when a manipulator uses actions other than trading to change the price of a stock. A classic example of this type of manipulation is the Harlem railway corner. In this case, the New York council passed a decree allowing the railway company to build a streetcar in certain areas of New York City. This increased the price of the stock greatly. Some of the council aimed to profit by selling this stock short and revoking the decree. This forces the price of the stock down by an action other than trading. Another example is when manipulators play with the value of their own company stocks. The managers of American Steel and Wire Company shorted their stock and then closed the company's steel mills. After announcing the closure, the stock price fell. The managers then covered their shorts and reopened the mills causing the price to rise again. These two examples are just some of the ways manipulators can impact a prices through actions other than trading.

1.3 Problem Definition

An example of an existing market manipulation approach is a top-down approach. This is done based on defined thresholds and patterns. Stock market data (such as price and volume) are monitored. By using a set of rules and red-flag triggers, possible fraudulent transactions are detected and further investigated. This method has its downfalls. It requires expert knowledge and would not be able to detect abnormal behavior from new and unknown manipulation schemes. This method does not adapt to the changing nature of the market and the exponentially growing amount of transactional data. Technology needs to evolve with this growth of data, and a better detection technique would make detecting manipulations much easier and efficient. Anomalies in stock market data are difficult to label as well. It is a long and tedious task and normal fluctuations in data might be mistakenly taken as anomalies. That is why finding a method that detects anomalies without the need to rely on labeled data is an interesting endeavor.

In this thesis, market manipulation will be detected through local anomaly detection using the behavior of similar time series. Companies within the same sector act in a similar way. They are generally affected by the same factors such as commodity prices, political or economical change. This proportionality can be exploited to detect market manipulation. Manipulations in stock data are represented as anomalies, where legitimate trades are normal behavior. One company's price can be related to that of similarly behaving companies to detect anomalies on a local level. Local anomalies are different from global anomalies. In local anomaly detection, a data point is considered anomalous with respect to its neighboring points. A local anomaly might not be considered anomalous when compared to all the other data points. Local anomaly detection is especially useful for non-

homogenous datasets and datasets with ever changing underlying factors, such as financial data.

Stock market trends are an ever evolving phenomena. The movement of the market today is very different than what it was a few years ago. That is because the things that effect the market are always changing. This indicates that to detect an anomaly in a given company, looking at its far past trends might not yield the best results since the behavior of the company and market have changed. Instead, one should look at similar companies and their behavior. This method of anomaly detection evolves with the stock market trends. It can also distinguish between market manipulations and normal, but unexpected, stock market spikes. For example, looking at the events following the UK's decision to leave the EU, the market crashed. If we are simply relying on a company's history, this can appear as an anomaly in the company's stock data because there was a sudden and sharp drop in price. However, this anomaly is not market manipulation, it is simply the markets reaction to a political situation. Using the proposed method, the company's price will be related back to similar companies. So external events that are not market manipulation would affect all the companies, and would not be considered anomalies. Market manipulation caused by individuals that target a company will be identified because it targets that single company.

1.4 Scope and Objectives

The aim of this thesis is to add on to the work of Golmohammadi *et al*. [1] who attempt to identify market manipulations using similar time series. Golmohammadi *et al*.

propose a prediction based Contextual Anomaly Detector (CAD). Their system starts by selecting a subset of time-series within the same sector of industry. Using this subset of time series, a centroid can be calculated by averaging all the time-series in this subset. The correlation between a given time-series and the centroid is then calculated, and a predicted value for that same time series is found. They define an anomaly as any point where the actual value deviates from the predicted value by more than a certain threshold. This method achieved an increase in recall when compared with random walk and k-nearest neighbor without compromising precision.

CAD provides the basis of an interesting solution to the problem. This thesis hopes to add onto this method by proposing a new method of preprocessing the data that can improve the method's recall without sacrificing precision. By using the Simple Moving Averages (SMA) of the companies' price changes, better recall can be achieved. This change in the way the data is preprocessed will be tested and compared with the original implementation of CAD and the k-means clustering algorithm.

The proposed solution will be tested by starting with clean manipulation free stock market data. Anomalies will then be inserted into clean data for the purpose of system evaluation. After that, the data will be preprocessed for normalization. This is where the new SMA step will be added. Then anomalies will be detected through the use of the centroid and the proposed prediction method. After that, the system will return possible manipulation instances for further inspection. Figure 1 shows an overview of this solution.

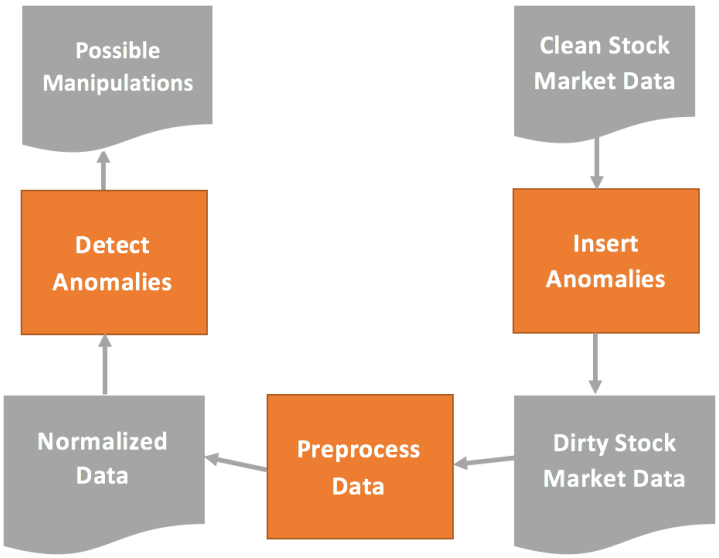

Figure 1: Solution Overview

1.5 Achievements

The thesis' goal is to improve the detection system's recall without hurting precision and this has been achieved by introducing a new preprocessing step. Using SMA to preprocess the time series data was more successful at finding manipulations. After performing experiments on two stock exchanges, the S&P and the Qatar stock exchange, recall was shown to undergo an improvement. For example, the recall of the consumer staples weekly dataset changed from 34% to 95%. Precision also saw an increase from 0.33% to 3.8%. This shows that the inherent similarities between sector companies can be exploited to monitor manipulations in an unsupervised manner.

1.6 Overview of Thesis

The thesis is divided into six chapters. The first chapter introduces the problem and gives some background information. Chapter two reviews related work in the field of market and securities manipulation. It summarizes some techniques that have been used to detect manipulations in this field and in time series data. After that, chapter three gives a detailed look into the proposed solution. It starts by describing the data collection process, how data is preprocessed, anomaly detection method and finally how anomalies are inserted for evaluation. Chapter four describes the experiment setup. Chapter five gives the experiment results and interprets them. Finally, the conclusion and future work are presented in chapter six.

CHAPTER 2: RELATED WORK

Anomaly detection techniques can generally be divided into supervised and unsupervised detection. Both of these techniques have been used to detect market manipulation and fraud. In this chapter, these methods will be reviewed.

2.1 Supervised Fraud Detection

Supervised anomaly detection requires the availability of labeled training data. These labelled instances are then used to build a predictive model that classifies normal and anomalous activities. New unlabeled data is compared through the built model to determine which class it belongs to. The problem with this methodology is that is it very difficult to obtain an accurate and representative labeled dataset. Another issue is that anomalous data points are much fewer in comparison to normal data points in the training datasets. This imbalance has to be addressed.

One such example of supervised learning used to detect market manipulation is the work of Ogut *et al*. [2] Market manipulation instances were labeled using cases published in the Capital Markets Board of Turkey. The aim of their research is to use two data mining techniques, Support Vector Machine (SVM) and Artificial Neural Networks (ANN) to build a predictive model and compare it with other statistical methods. Their experimental results show that SVM and ANN outperform the statistical techniques.

Diaz *et al*. [3] also constructed a labeled dataset using manipulation case studies from the US Securities and Exchange Commission during 2009. Their work detected

intraday price manipulation by using financial variables and ratios along with textual sources. They applied tree generating learning methods, QUEST, C5.0 and C&RT.

Golmohammadi *et al*. [4] used the data supplied in the works of Diaz et al. [3] and expanded on it to explore different classifiers for market manipulation detection. They applied the following classification methods: CART, inference trees, C5.0, random forest, naïve bayes, neural networks, SVM and k-nearest neighbor. Their experimental results show Naïve Bayes outperforms all other classifiers achieving an F2 measure of 53%.

2.2 Unsupervised Fraud Detection

Unsupervised anomaly detection does not require any labelled data, and is thus the most widely applicable. This technique assumes that normal data points are much more frequent than anomalies. This is the more flexible technique when compared to supervised methods.

2.2.1 Regression

Many economic models intend to forecast stock market trends and detect manipulations through linear regression, auto-regression moving average (ARMA), autoregressive integrated moving average (ARIMA), and other such models [5]–[7]. However, these statistical models can only handle linear data. They do not handle highly noisy, irregular, non-linear data such as stock market data. These approaches often fail to predict the market and manipulations properly.

Yang *et al*. [8] used logistic regression models to detect market manipulation in

the Shanghai and Shenzhen market. Market characteristics are analyzed using primary component analysis to increased the model's forecasting performance. The model proved better than linear regression models with a higher success rate for prediction.

2.2.2 Rule Induction

Rule induction is a data mining technique that has been used to detect fraud in the stock market. This method draws similarities to existing regulatory rules that are used to monitor the market, this makes it very popular among auditors and securities market investigators. In Jungwon *et al* [9], a fraud detection method is implemented using rule induction. First, they randomly generate rules using the association rules algorithm *Apriori*. Then, they apply these rules to a dataset containing only legitimate transactions. Any rule that matches the data is then disregarded. Then, the remaining rules are used to monitor new transaction data in the system. Any rule that detects an anomaly is replicated by adding a small random mutation. All successful rules are retained for anomaly detection.

Associative rules were also used to detect manipulations in intraday trades in the Thai bond market in the work of Mongkolnavin *et al.* [10] Price variations and investor behavior were intergrated to analyze red-flag signals in real time. A single trade transaction involves the trader ID, and the transaction order, where the last transaction is assigned zero, the second to last is assigned one and so-on. The system detects possible manipulations if an association between trader ID and transaction order is found. In normal cases, no association should be present. Trading time should be random. If a high association is present, then that trader is a suspected manipulator.

2.2.3 Hidden Markov Model

Hidden Markov Model (HMM) assumes that the underlying process of the time series is a hidden markovian process. The time series training data is used to build an HMM that assigns anomaly scores to the test time series. It is used in anomaly detection in intrusions detection systems as in the work of Jecheva [11]. This method assumes that there is a hidden markovian process that generates the normal time series.

HMM is customized for market manipulation detection in the work of Cao *et al*. [12] A traditional retraining mechanism is applied to automatically track the changes in the statistical properties of the time series. The proposed Adaptive Hidden Markov Model with Anomaly States (AHMMAS) was tested on US and UK market data. Its performance out did the k-nearest neighbors algorithm, gaussian mixture models and one-class support vector machines.

2.2.4 Visualization

Instead of looking at stock market data as independent data points, it is also interesting to consider the relationship between sellers and buyers. Using these relationships, a social network can be composed where nodes represent entities or objects and edges can be dependencies or relationships. Blume *et al*. [13] combined Social Network Analysis (SNA) and interactive visualization to identify fraudulent accounts in an exchange. They defined indicators of fraud based on textual descriptions of fraud cases to identify these accounts. Using SNA can identify circular trading which is when an account is consistently buying and selling the same volume of stock. It can also flag when an account buys low and sells high. SNA has many useful algorithms that are applicable to stock data for the sake of fraud detection.

Financial investigators and regulators are always referring to charts and figures when monitoring the market. Visually delivering complex information and patterns within the data is of great interest. Huang *et al*. [14] deliver a visual analytic framework for stock market security. It consists of two parts, the first handles visual surveillance of market performance through the use of a 3D treemap. Each cell in the 3D visual represents a security, the size represents volume and the color indicates whether a change is an increase (green) or decrease (red). The treemap provides a tool for real-time data visualization. The trading data is compared to a set of parameters and an alert is flagged when the data is out of that range. The second part visualizes trading networks for SNA and monitoring broker's activities. Nodes represents traders, the directed edges represent the flow and weight of trades. To identify suspicious trades, a database of past malicious trading patterns is referenced.

2.2.5 Anomaly Detection

Anomaly or outlier detection techniques look for data that appear inconsistent with most of the data observations. Ferdousi *et al*. [15] applied such a technique to transactional data to find outliers. They used Peer Group Analysis (PGA) on three months of Bangladesh stock market data. The data consists of statistical variables such as mean and variance along with buy and sell orders. The objective of PGA is to categorize target objects into peer groups. This is decided by looking for the peer group that is made up of the most similar objects to the target object. After a certain timeframe, five weeks in the experiment, a centroid is found for each peer group. After that, the distance of each group member to the centroid is calculated using t-statistics. Objects that are significantly far from their peer

centroid are then flagged as outliers. Traders associated with that object are then put under suspicion of fraud since they behaved differently than their peers.

Vlachos *et al*. [16] employ burst event detection for correlating surprising volume trading events in the New York stock exchange. Bursts are identified in the data based on a variable threshold using the skewed nature of financial data. The bursts are then indexed for efficient access using Containment Encoded Intervals (CEIs). Correlated bursts are identified by performing overlap operations on the indexed burst regions. This method can be used to identify fraud in real-time due to the proposed indexing technique. This approach was tested using historical trading data before and after the events of 9/11 with results showing the method is superior to B+ tree.

Golmohammadi *et al*. [1] expanded on their previous work in this field by introducing an unsupervised approach for market manipulation. The authors developed a prediction based Contextual Anomaly Detector (CAD). This method exploits the fact that stocks within the same sector behaving and react similarly. The proposed method works for complex time series that do not follow a deterministic model such as stock markets. First, they take a subset of time-series in a given sector based on a window size. Then, a centroid is calculated by taking the mean of the data points in the time-series subset at every time point. This centroid represents the expected behavior of the subset. The centroid along with the Pearson correlation of the time series with the centroid is used to get a predicted value of a stock within the subset. This predicted value is compared with the actual value of the stock using the Euclidean distance to get an anomaly score. If the anomaly score is greater than a certain threshold (the standard deviation of the series is

used), then this point is an anomaly. The method improves recall from 7% to 33% when compared with k-nearest neighbor and random walk without compromising precision.

2.3 Discussion

By reviewing the related work, a history of the field was established. Prediction methods that rely on linear regression were not the best at solving this problem due to the stock markets complex nature. That is why the work that forecasts prices of a company using similar time series is interesting. The method is a recent contribution that can be further investigated.

CHAPTER 3: METHODOLOGY

In this approach, the behavior of similar time-series is used to detected anomalies. If a company deviates far from how it is expected to behave given similar companies, an anomaly or market manipulation could be present. Before presenting how this is achieved, we must fully understand the data and prove if there is a similarity between companies within one sector. After that, we will tackle how the data is preprocessed and anomalies are identified. Lastly, this chapter will detail how anomalies are inserted for the purpose of evaluation.

3.1 Data Collection

The method is tested against two dataset collections. The first is the same data used in the works of Golmohammadi *et al.* [1]. The second is data from the Qatar Stock Market (QSM). All sets of data contain the time aspect as the first attribute, in both daily and weekly increments. The other attributes are the companies' names, with the closing price of that day or week.

The data used in [1] are divided into different industry sectors of the S&P 500 index. The S&P 500 is an American stock market index made up of the largest 500 company stocks listed on the NYSE or NASDAQ indices. It is one of the most followed indices and is considered one of the best representations of the US stock market. These are widely assumed to be manipulation free since they are highly liquid and closely monitored by regulatory organizations.

The second set of datasets have been extracted using the Bloomberg API.[3] The data belonging to Qatari companies are divided into sectors similar to the S&P 500 datasets.

3.2 Correlation Study

To determine whether the assumption that companies in the same sector behave similarly, a correlation study was conducted. This was done by calculating the correlation coefficient. In finance and investment, the correlation coefficient is a measure that represents whether two variables move in the same way. The values of the coefficient range from 1 to -1. A correlation of -1 means we have a perfect negative correlation, where the variables are inversely proportional. While a value of 1 is a perfect positive correlation, where variables are proportional. It can be calculated using the following formula:

$$\rho_{xy} = \frac{cov(r_x, r_y)}{\sigma_x \sigma_y} \quad (1)$$

where: ρ_{xy} is the Pearson product-moment correlation, $cov(r_x, r_y)$ is the covariance of the two series r_x and r_y under investigation, $\sigma_x$is the standard deviation of series r_x, and σ_y is the standard deviation of series r_y. The covariance measures how the two series change together, however, the magnitude is unbounded and difficult to interpret. Dividing the covariance by the product of the two standard deviations normalizes the value of the statistic. This makes it easier to interpret.

This statistic is useful in many financial applications. It can help determine how

[3] Bloomberg API, https://www.bloomberglabs.com/api/

well a mutual fund is doing compared to a certain benchmark, or to determine how two funds relate to each other. By adding a negatively correlated fund to their portfolio, investors can diversify their gains. Because this measure is so widely used by the financial world, it is what was used to determine whether time series in the same sector truly behave similarly.

Table 1 contains a sample of the correlation matrix of the S&P consumer staples dataset. The correlation coefficient between the companies is close to one. This confirms the similarity between them and this proportionality can be used to monitor any possible manipulation.

Table 1. S&P Consumer Staples Correlation Matrix

Company	Walmart	Walgreen	Coca Cola	Colgate-Palm	General Mills
Walmart	-	0.95	0.89	0.96	0.94
Walgreen	0.95	-	0.86	0.93	0.91
Coca Cola	0.89	0.86	-	0.90	0.90
Colgate-Palm	0.96	0.93	0.90	-	0.98
General Mills	0.94	0.91	0.90	0.98	-

The S&P index is made up of large well-established companies. Most sectors are made up of highly correlated companies. The only sector that does not follow this assumption is the Information Technology sector. Table 2 contains a sample of the IT sector correlations. This sector contains a more varied selection of companies such as Apple, Cisco and Visa. This variation yielded a much less correlated collection of

companies.

Table 2. S&P IT Correlation Matrix

Company	APPLE	GOOGLE 'A'	INTEL	FACEBOOK CLASS A	CISCO SYSTEMS	VISA 'A'
APPLE		0.80	0.32	-0.05	0.20	0.76
GOOGLE 'A'	0.80		0.35	0.86	0.28	0.96
INTEL	0.32	0.35		0.41	0.92	0.61
FACEBOOK CLASS A	-0.05	0.86	0.41		0.42	0.84
CISCO SYSTEMS	0.20	0.28	0.92	0.42		0.26
VISA 'A'	0.76	0.96	0.61	0.84	0.26	

Most sectors in the S&P exchange are well-established and thus contain highly correlated companies. However, does this correlation between sector companies hold for a newer and younger market such as the QSM. A similar study was conducted and it appears that QSM sectors are much less correlated. Some companies are almost completely uncorrelated. Table 3 shows the correlation matrix of some of the companies in the consumer goods and services sector in the QSM. This is because the QSM is made up of relatively younger companies. Their trends differ from one another due to their immaturity and other differences.

Table 3. QSM Consumer Goods and Services Correlation Matrix

Company	ZAD Holding	Q.German Medical	Salam International	Medicare	Qatar Cinema
ZAD Holding		0.58	0.59	0.93	-0.07
Q.German Medical	0.58		0.57	0.58	0.22
Salam International	0.59	0.57		0.60	0.44
Medicare	0.93	0.58	0.60		0.04
Qatar Cinema	-0.07	0.22	0.44	0.04	

3.3 Preprocessing

The times series data should be normalized before detecting the anomalies. This step is crucial as it is with many machine learning and statistical methods. The data contains closing prices of stocks which is the most important feature when monitoring the market for manipulation. However, this leaves a wide range of prices in the dataset. The actual price of the stock when related to other companies is not of interest here. What is important is how this company's price changes in relation to other companies. Similar companies undergo similar price changes even if their prices are very different.

The proposed preprocessing step in [1] is to use the price percentage change. This is done using the following formula:

$$Ch_t = \frac{(P_t - P_{t-1})}{P_{t-1}} \quad (2)$$

where Ch_t represents the change at time t, P_t is the price of the stock at time t, and $P_{t-1}$is the price at *t-1*. This normalizes the data and scales the stock prices. Every row is dependent on the row before, however, it does not contain any information reflecting the history of

the stock beyond that.

Changing how the data is preprocessed can improve the results. The data can be normalized while still reflecting the history of the stock price movement. This thesis proposes the use of simple moving average (SMA) instead. SMA can be calculated using the following formula:

$$SMA_t = \frac{(Ch_t + Ch_{t-1} + Ch_{t-2} + \cdots + Ch_0)}{n} \quad (3)$$

where SMA_t is the simple moving average at t, Ch_t is the price change of the stock at time t, $Ch_{t-1}$is the price change at t-1, $Ch_0$is the initial price change at the first instance in the dataset, and n is the number of instances from the start of the dataset to instance t. This formula normalizes and scales the data while still maintaining information regarding the stocks history. This new prepressing step improves recall greatly. The two method will be tested against one another in chapter 4.

3.4 Anomaly Detection

Normally, anomaly detection involves comparing a new sample with a given set of normal samples. By measuring these two against one another, an anomaly score is assigned to the new sample. High scoring samples would then be labeled as anomalies. In a field as complicated as finance, distinguishing between normal and abnormal behavior is not so black and white.

Using supervised learning for fraud detection in stock market has yielded good results. However, obtaining an accurately labeled dataset is difficult and impractical. Labeling such datasets involve going through litigation cases and taking these observations

as anomalies. The rest of the data would be then labeled as normal. The other way is to generate a synthetic dataset. However, learning from synthetic data might not be very reflective of how manipulations are in real life.

Developing an unsupervised detection method for market manipulation eliminates the costs of manually labelling data. A solution is presented in the Contextual Anomaly Detection (CAD) method [1]. The CAD system proposes detecting manipulation through prediction. Unlike other prediction based systems, this does not assume the series is following a deterministic model. Instead of relying on the series historical data for prediction, the behavior of similar series is used to predict the series next value.

Golmohammadi *et al*. develop CAD to take a set of time series from one sector and a window size. It first takes a subset of time series based on the window size. Then, a centroid is calculated by taking the mean of the time series at every time instance *t*. This centroid is a representative of the expected behavior of this subset. Figure 2 illustrates how the centroid of the S&P energy sector would look like given a set of company time series. This centroid is used along with a feature of each time series to get a predicted value for that time series.

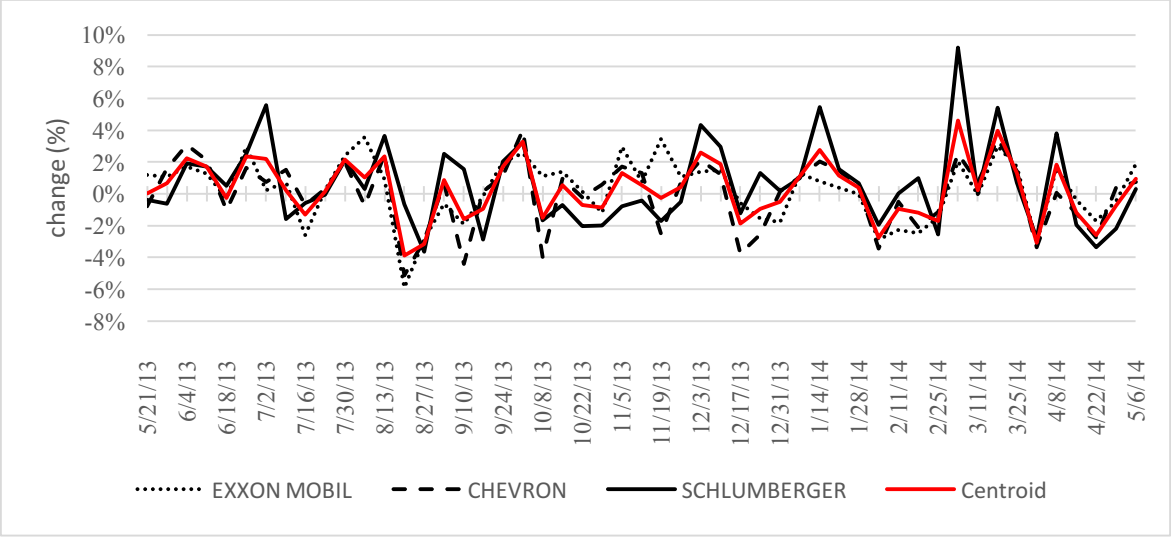

Figure 2. Centroid Time Series of Stocks in S&P 500 Energy Sector

In more detailed terms, the system would start with a subset of similar time series $\{X_i | \, i \in \{1,2,\dots,d\}\}$. After that a centroid is found by taking the average of the time series values at each time *t*. The centroid series is represented as $\{C | \, i \in \{1,2,\dots,n\}\}$, where n is the last day or week in the time series. The predicted value of a time series X_i at time t is then calculated as:

$$\hat{x}_{it} = x_{it-1} * cor(X_i, C) \tag{4}$$

where x_{it-1} is the value of series X_i at *t-1* and $cor(X_i, C)$ is the Pearson correlation of X_i and the centroid *C*. This correlation is used because since the centroid is a representative of how the collection of series moves, its correlation with a certain time series can help predict its value. The predicted value $\hat{x}_{it}$ is then compared with the actual value of X_i at time *t* by using the Euclidean distance:

$$\epsilon = \sqrt[2]{(\hat{x}_{it} - x_t)^2} \quad (5)$$

ϵ is the anomaly score that is then compared to the standard deviation of the company under investigation. If the anomaly score is greater than the standard deviation, x_t is an anomaly.

The proposed CAD system uses the percentage change value to represent the time series in the data. Using this method, the value of x_{t-1} is used to get the prediction of the next value in the series. However, these values are only representative of the value right before it, since they are calculated by comparing two consecutive rows. They do not contain any information relevant to the history of the data. That is why the new method of preprocessing is proposed. Through using the SMA_t, we are using a more meaningful value that holds some information regarding the change pattern of the company.

CAD is a local anomaly detection method, it works by taking window size that is used to get a subset of companies in a sector and by using certain periods of data (for example, a year or two).

3.5 Anomaly Insertion

For the purpose of evaluation, synthesized anomalies will be inserted into the datasets. The datasets used are regarded as manipulation free, because the S&P is made up of heavily regulated companies. They are also some of the largest companies in the US and highly liquid. Highly liquid companies have buyers and sellers trading at all times so its very difficult for one party to affect the stock price or take control. This makes this dataset ideal for testing forecasting or fraud detection and has been used in many research studies [1], [2], [17], [18].

In a collection of normal stock behavior, manipulation would represent itself as an anomaly or outlier. Market manipulation can appear as any abnormal jump or drop in a company's price, these outliers would represent possible manipulations. By using a well-known definition of outliers, manipulation cases will be added to the manipulation-free data for evaluation. For the sake of comparison, the same injection method used in the work of Golmohammadi *et al*. is applied. They adopted Turkey's method for time series that do not follow a normal distribution [19]. In a normal distribution, an outlier is defined as an instance that is three standard deviations away from the mean. However, in distributions with skewness outliers should be defined differently. Turkey's method defines outliers according to the following formula:

$$O(x_{it}) = \begin{cases} \mu + (Q_3 + (3 * IQR)\ if\ \gamma > \epsilon,\ upperbound \\ \mu - (Q_1 - (3 * IQR)\ if\ \gamma > \epsilon,\ lowerbound \\ \mu \pm 3\sigma\ if\ \gamma = 0, \qquad if\ normal\ distribution \end{cases} \tag{6}$$

$O(x_{it})$ is the outlier specific for the time series X_i, Q_1 is the 25th percentile (1st quartile), Q_3 is the 75th percentile (3rd quartile), IQR is the inter-quartile range $(Q_3 - Q_1)$, γ is the skewness, μ is the mean and σ is the standard deviation. With series with skewed distributions, the first step in defining an outlier is finding the statistical center of the range. This is done with the use of the 1st and 3rd quartiles. A quartile is a division of the data into four sectors based on the data values in the series. The skewness of the series is found using:

$$\gamma = \frac{\mu_3}{\mu_2^{\frac{3}{2}}} \tag{7}$$

μ_2 and μ_3 are the second and third central moments or moments about the mean.

It is important to note that outliers are defined according the the time series itself

independent from the other series in the dataset. The detection method is also unrelated to the synthetic outliers. Supervised methods would learn from the training data so using synthetic outliers in those circumstances would highly affect the end result. In this system they are merely used for the purpose of evaluation.

3.6 Discussion

The methodology behind CAD takes advantage of the correlation between sector companies to find anomalies. Stock market data has very different prices across different companies. Finding a suitable normalization method is essential. Using SMA to normalize data gives a more representative value for the time series. This will be tested with the proposed anomaly insertion method in the following chapters.

CHAPTER 4: EXPERIMENT SETUP

To fairly compare CAD with SMA, the experiment setup should try to replicate the original experiment closely. The data of the sector companies will be tested in both daily and weekly increments and a variety of experiments will be conducted. The following chapter will detail these experiments and how the system will be evaluated.

4.1 Data

As was mentioned in chapter 3, two data collection are used to evaluate the system. The S&P dataset is divided into five sectors: consumer staples, consumer discretionary, energy, finance and technology. Each sector contains the time series of some of the largest US companies in its respective sectors. These datasets are considered manipulation free with no anomalies.

The second dataset is related to local Qatari companies. They are also divided into sectors: finance, consumer goods and services, industrial and insurance. The method will be tested on the new Qatar dataset to see whether the CAD system will be effective with a much smaller and younger market. The Qatari stock markets is made up of only 43 companies as opposed to the 500 in the S&P index. Only four of the largest sectors of the QSM will be studied because the rest are too small and only contain four or less companies. Table 4 details the sizes of the datasets. They will be used in both daily and weekly increments.

Table 4. Dataset Used for Experiments

S&P 500			
Sectors	Number of Companies (time series)	Number of weekly instances	Number of daily instances
Consumer Staples	40	64,000+	323,000+
Consumer Discretionary	85	111,000+	558,000+
Energy	44	64,000+	315,000+
Financial	83	117,000+	587,000+
Information Technology	66	80,000+	395,000+
Qatar Stock Exchange			
Financial	13	700+	5,000+
Consumer Goods & Services	9	700+	5,000+
Industrial	8	700+	5,000+
Insurance	5	700+	5,000+

4.2 Baseline

The aim of this thesis is to evaluate the new preprocessing step against the one used in the original CAD system. The following experiments will be designed to replicate the one in the original work as much as possible to ensure a fair comparison. CAD was tested using multiple window sizes (15, 20, 24) on the five sectors in both daily and weekly increments.

4.3 Experiments

The experiments were set up to replicate the original CAD experiments. The same

window sizes are used (15, 20, 24) over multiple periods of time (one, two and four years). The percentage of outliers injected into the datasets was equal to 0.1% of the total number of data points in the sets. This is set low to reflect how manipulations would present themselves in real life. This will be used with the original preprocessing step and the new preprocessing step to compare their performances. After that, the percentage of anomalies will be increased to test its effect on the system's performance.

Another well know unsupervised learning method will be tested. K-means clustering is a widely used clustering algorithm and is very efficient with respect to execution time. This method aims to partition the observations into k clusters in a way where each observation belongs to the cluster with the nearest mean. It starts by first defining k centroids, one for each cluster. The centroids should be placed well, since different starts could result in different results. The usual practice is to place the k centroids as far away from each other. Then each data point is assigned to the nearest centroid. When all the points are distributed, new k centroids are calculated as bar centers of the clusters. The data are again redistributed according to these new centroids. This is repeated until the reassignments cause no change. Centroid based cluster methods have shown to be better than density based clustering methods for financial data. [20] This method will attempt to cluster the stock data into two clusters, abnormal and normal, and its performance will be measured.

4.4 Evaluation metrics

To evaluate the system, the measures *recall, precision* and *F-measure* will be used.

Recall is the fraction of manipulations that are retrieved by the system, and precision is the fraction of actual manipulations in the retrieved data. Manipulations in such regulated environments are not very frequent, however, missing any manipulation could cost companies a lot of money. In these circumstances, the cost of misclassification is not equal. That is why the aim of the experiments is to achieve high recall. A false negative can cause companies much more than a false positive. Although, avoiding false positives is very desired, it is not the focus of the work. The aim is to improve recall while not sacrificing precision. For that reason, a higher value of β for F-measure is used to give more priority for recall than precision.

4.5 Discussion

These experiments are designed to effectively evaluate the new preprocessing step against the original system. They are also intended to replicate how market manipulation will be handled in a real world scenario. Anomaly percentage is set low to mimic how uncommon these manipulations are. Recall, precision and F-measure will reflect system performance and give a sense of how well the system is at detecting these manipulations.

CHAPTER 5: RESULTS

In this chapter the experiment results are presented. First, the CAD will be compared with CAD-SMA that has the new preprocessing implementation. These will also be compared with the results of the unsupervised clustering method, simple k-means. After that, the effects of varying the amount of anomalies in the data will be studied. Finally, the method will be implemented to detect anomalies in the QSM data.

5.1 Comparison of Algorithms

After running the experiments, the new preprocessing step proved to improve recall significantly. Table 5 shows the performance of CAD as it was proposed in the original work, compared with CAD with the SMA preprocessing step and the simple k-means clustering algorithm for the weekly S&P data (Table A.1 in the Appendix shows the complete comparison of both weekly and daily data). The bellow performance was achieved using a window size of 15 and with 0.1% anomalies. The results are representative of all window sizes (15, 20, 24), since the methods return very close measures regardless. These measures are also stable regarding the number of time instances. The method was tested with periods of 4 years, 2 years and 1 year.

Table 5. Comparison of CAD, CAD–SMA and Simple K-means Applied on Weekly S&P Data

S&P 500					
Dataset	Algorithm	Recall (%)	Precision (%)	F2 (%)	F4 (%)
Consumer Staples Weekly	CAD	34.7	0.33	1.59	4.86
	CAD - SMA	95.05	4.7	19.62	44.61
	Simple K-means	40.01	3.8	13.77	25.64
Consumer Discretionary Weekly	CAD	34.15	0.33	1.6	4.88
	CAD - SMA	88.03	1.5	7.02	20.04
	Simple K-means	40.7	3.6	13.30	25.34
Energy Weekly	CAD	34.49	0.33	1.58	4.83
	CAD - SMA	86.7	7.4	27.58	53.18
	Simple K-means	40.2	1.5	6.53	15.97
Financial Weekly	CAD	35.47	0.34	1.65	5.05
	CAD - SMA	97.23	2.9	12.95	33.37
	Simple K-means	40.74	3.62	13.35	25.41
IT Weekly	CAD	33.69	0.34	1.63	4.98
	CAD - SMA	96.41	6.92	26.88	54.76
	Simple K-means	18.32	0.83	3.51	8.18

The results show that the new preprocessing step improved both recall and precision as compared with the other two algorithms. Recall is what is more important in a problem such as market manipulation detection. Almost all anomalies were retrieved, however, precision is very low. It is slightly better than CAD precision, but still there is a great chance of improvement. F2 and F4 measures are used to give recall a higher weight in the harmonic mean. This is because the cost of false negatives and false positives are not equal. False positives are undesirable, but missing manipulation can be very expensive for

companies.

When using SMA, the value at *t-1* carries more information regarding the historical movement of the time series. So when predicting the value of the time series at *t,* a more accurate prediction is calculated and it produces better recall and precision measures. However, by using SMA as the normalization method, the anomalous value gets propagated into the following data in the series. In the original preprocessing method, the anomalous entry is only reflected in the one entry and affect the value immediately after it. The anomalous jump in price is no longer carried into the other series entries. SMA averages the changes over time so it carries the anomaly over into the subsequent entries. This should explain the low precision value. More normal entries are labeled as abnormal because of how the anomalies effect is still carried in the data from one row to the other.

A solution to this problem would be to calculate the SMA of an entry for over a certain number of rows. For example, calculate the SMA at *t* for the change percentage from *(t-10)* to *t*. This way the effect of the anomaly will fade over the next ten rows. This can be combined with an anomaly threshold. An anomaly threshold can be the effect radius of an anomaly. When an anomaly is detected, the threshold can define the number of rows where the anomalies effect might still be present. Another solution would be to use the Exponential Moving Average (EMA) is a type moving average where more weight is assigned to the latest data. This will also dilute the effects of the anomaly over time.

The simple k-means clustering algorithm achieved better recall and precision than the CAD algorithm in almost all the sectors. The one sector that hasn't performed as well was the information technology sector (IT). This could be because it is in its nature a more

erratic sector. IT companies in the S&P are very varied and range from telecommunication companies such as Broadcom Ltd to the gaming company Activsion Blizzard Inc. Table 2 in the correlation study showed a sample of how less correlated this sector is compared with the other sectors. Clustering the data in this sector did not yield the same results as with other sectors. K-means clustering works best with data that can be spherically clustered. Centroid based clustering such as k-means preform better than density based clustering, however, it does not handle non-global data such as financial data as well as other data [20]. This might be most apparent in a diverse sector such as IT.

5.2 Varying Anomaly Percentage

The experiments in Table 5 were applied to a dataset with 0.1% anomalies insertion. Table 6 shows how CAD-SMA behaves with increasing the number of anomalies. The percentage of anomalies inserted will be varied on the consumer discretionary weekly dataset.

Table 6. CAD-SMA on Consumer Discretionary Weekly with Varied Anomaly Percentage

Anomaly percentage	Recall	Precision	F2	F4
0.10%	95.05	4.7	19.62	44.61
5%	68.66	73.27	69.54	44.61
10%	54.64	77.86	58.11	68.92
25%	42.44	84.5	47.13	55.62

As the number of anomalies increase, the recall decreases. On the other hand, precision is improving. CAD relies on the assumption that series are highly correlated, and thus one series' behavior can be used to predict another. However, as the number of anomalies increase the recall decrease. This could be because as more anomalous values are inserted into the series, the less their behavior as a collective group is correlated. These new entries will cause the series to deviate from their expected behavior and make them less similar or correlated. This might be why the method is less effective with higher anomaly percentages.

5.3 Detecting Anomalies on QSM Dataset

Figure 3 shows the recall and F4-measure of CAD, CAD-SMA and simple K-means on the weekly QSM sectors. Best Recall is achieved by using CAD-SMA. Using this algorithm, average recall is high and most anomalies are retrieved. On the other hand, F4 measure is lower than those of the S&P sectors using all algorithms. That is because precision of these methods on Qatari data were lower than that of S&P (full measures for both weekly and daily sets are in table A.2 in the Appendix).

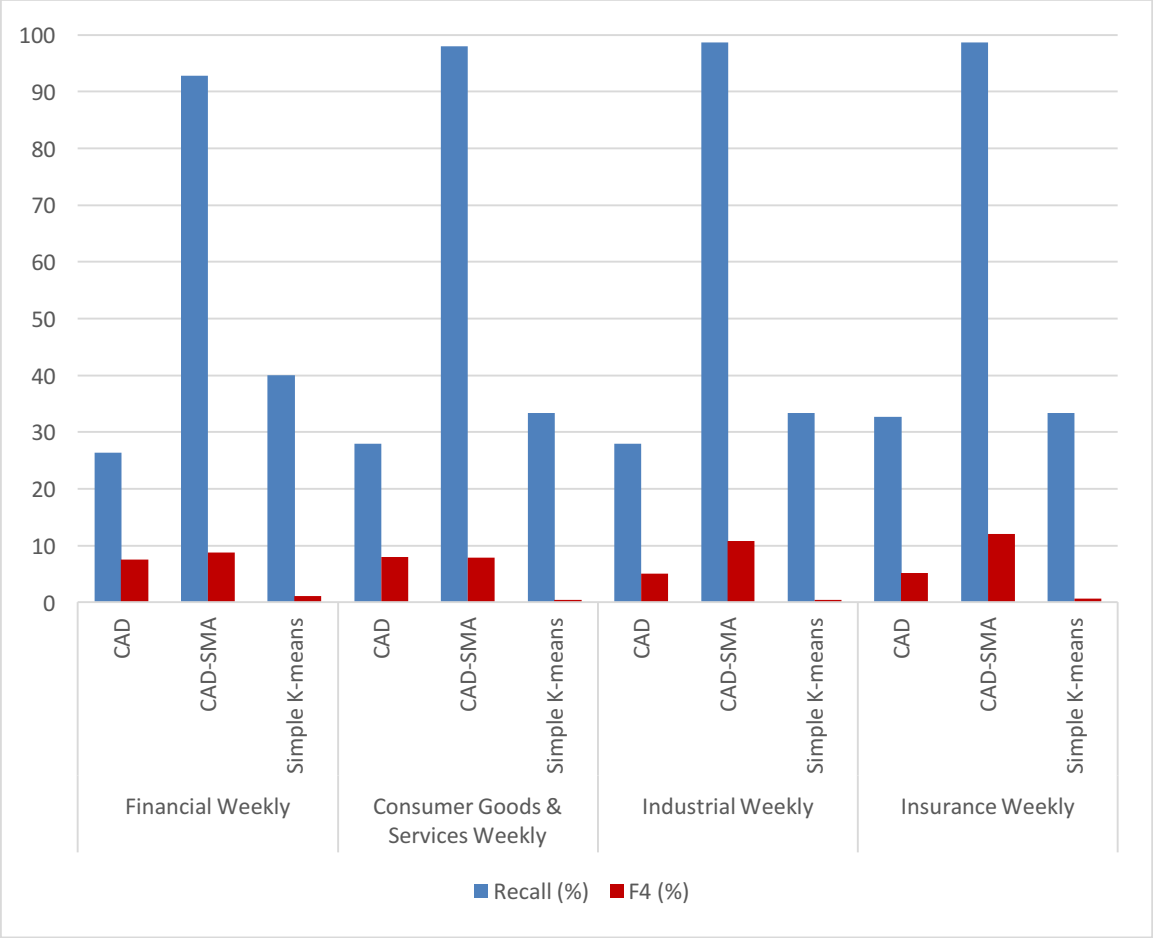

Figure 3: Comparison of Recall and F4 measure using CAD, CAD-SMA and K-Means on weekly QSM data

The low precision could be due to the size of the QSM. This market is much smaller in size than the S&P. The companies in QSM are also much younger. Young companies tend to behave very differently than older, well-established companies. This causes some of companies in each sector to be less correlated than the S&P sector. These factors could be attributed to why the method is less effective with the QSM dataset.

5.4 Discussion

These experiments provide valuable insights into the strengths and weaknesses of

these methodologies. By analyzing these results, we can conclude when this method is best used and the areas that can benefit from further innovation. The precision of this method is not as good as it should be. The manipulations are detected but many other normal behaviors are falsely reported as well. This can be further improved in order to improve the detection system.

CHAPTER 6: CONCLUSION

Unsupervised anomaly detection is a fascinating field that has a lot yet to be discovered. Stock market manipulation detection is a unique subset in this field. Stock market time series has many important properties that need to be taken into consideration when tackling this problem. Instead of using regression models or other pre-existing forecasting models, one can look at similar time series. The behavior of similar time series in market sectors can be used to monitor a company that can fall victim to manipulation.

6.1 Summary

Through using the prediction-based anomaly detector CAD, market manipulations were detected. A centroid was found by averaging the collection of similar time series. Then anomalies are found by using the correlation of the centroid with time series. Before that, the prices in the stock data need to be scaled and normalized.

When dealing with market data, normalization is crucial for best results. Actual stock prices are not as interesting as the stock price changes. The CAD algorithm suggested the use of percentage change from one price to the other. However, using the simple moving average of the percentage change has proven more meaningful. That is because it is more reflective of the history of the time series. This improved recall, however, precision was still low.

6.2 Future Work

The detection method had very low values for precision. A solution to this would be to restrict the SMA calculation for a certain number of rows. This will make the effect of the anomaly fade over a certain number of row. If this is used along with a threshold for the anomaly effect, this could improve precision. Another solution would be to use the Exponential Moving Average (EMA). This give a higher weight to more recent changes in the prices. Using this will also lessen the effect of the anomaly after a certain number of rows. There's still room for innovation in this method to improve precision. Unsupervised anomaly detection is a challenging field, but its applications are very rewarding.

REFERENCES

[1] K. Golmohammadi and O. Zaiane, "Time Series Contextual Anomaly Detection for Detecting Market Manipulation in Stock Market," *Data Science and Advanced Analytics (DSAA)*, 2015.

[2] H. Öğüt, M. Mete Doğanay, and R. Aktaş, "Detecting stock-price manipulation in an emerging market: The case of Turkey," *Expert Systems with Applications*, vol. 36, no. 9, pp. 11944–11949, 2009.

[3] D. Diaz, B. Theodoulidis, and P. Sampaio, "Analysis of stock market manipulations using knowledge discovery techniques applied to intraday trade prices," *Expert Systems with Applications*, vol. 38, no. 10, pp. 12757–12771, 2011.

[4] K. Golmohammadi, O. R. Zaiane, and D. Díaz, "Detecting Stock Market Manipulation using Supervised Learning Algorithms," 2011.

[5] S. Radha and M. Thenmozhi, "Forecasting short term interest rates using ARMA, ARMA-GARCH and ARMA-EGARCH models," in *Indian Institute of Capital Markets 9th Capital Markets Conference Paper*, 2006, pp. 1–14.

[6] R. Ballini, I. Luna, L. M. De Lima, R. Lanna, and F. Silveira, "A comparative analysis of neurofuzzy , ANN and ARIMA models for Brazilian stock index forecasting," *SCE-Computing in Economics and Finance*, 1995.

[7] A.Angabini & S. Wasiuzzaman, "GARCH Models and the Financial Crisis-A Study of the Malaysian Stock Market," *The International Journal of Applied Economics & Finance*, vol. 5, no. 3. pp. 226–236, 2011.

[8] F. Yang, H. Yang, and M. Yang, "Discrimination of China's stock price

manipulation based on primary component analysis," *Proceedings of 2014 IEEE International Conference on Behavioral, Economic, Socio-Cultural Computing, BESC 2014*, no. 11, 2014.

[9] K. Jungwon, A. Ong, and R. E. Overill, "Design of an artificial immune system as a novel anomaly detector for combating financial fraud in the retail sector," *2003 Congress on Evolutionary Computation, CEC 2003 - Proceedings*, vol. 1, pp. 405–412, 2003.

[10] J. Mongkolnavin and S. Tirapat, "Marking the Close analysis in Thai Bond Market Surveillance using association rules," *Expert Systems with Applications*, vol. 36, no. 4, pp. 8523–8527, 2009.

[11] V. Jecheva, "About Some Applications of Hidden Markov Model in Intrusion Detection Systems," *International Conference on Computer Systems and Technologies*, pp. 1–6, 2006.

[12] Y. Cao, Y. Li, S. Coleman, A. Belatreche, and T. M. McGinnity, "Adaptive hidden Markov model with anomaly states for price manipulation detection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 2, pp. 318–330, 2015.

[13] M. Blume, C. Weinhardt, and S. Detlef, "Using network analysis for fraud detection in electronic markets," in *Information Management and Market Engineering*, 4th ed., KIT Scientific Publishing, 2006, pp. 101–112.

[14] M. L. Huang, J. Liang, and Q. V. Nguyen, "A visualization approach for frauds detection in financial market," *Proceedings of the International Conference on Information Visualisation*, pp. 197–202, 2009.

[15] Z. Ferdousi and A. Maeda, “Unsupervised Outlier Detection in Time Series Data,” *22nd International Conference on Data Engineering Workshops ICDEW06*, pp. x121–x121, 2006.

[16] M. Vlachos, K.-L. Wu, S.-K. Chen, and P. S. Yu, “Correlating burst events on streaming stock market data." 16.1 (2008): 109-133.,” *Data Mining and Knowledge Discovery*, vol. 16, no. 1, pp. 109–133, 2008.

[17] W. Huang, Y. Nakamoria, and S.-Y. Wang, “Forecasting stock market movement direction with support vector machine,” *Computers and Operations Research*, vol. 32, no. 10, pp. 2513–2522, 2005.

[18] D. Enke and S. Thawornwong, “The use of data mining and neural networks for forecasting stock market returns,” *Expert Systems with Applications*, vol. 29, no. 4, pp. 927–940, 2005.

[19] J. W. Turkey, *Exploratory Data Analysis*, 18th ed., vol. 2. Addison-Wesley Publishing Company, 1977.

[20] F. Cai, N.-A. Le-Khac, and T. Kechadi, “Clustering Approaches for Financial Data Analysis: a Survey,” in *International Conference on Data Mining (DMIN 2012)*, 2012.

[21] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly Detection : A Survey,” no. September, pp. 1–72, 2009.

APPENDIX A: TABLES

Table A.1. Complete Results of S&P Datasets with CAD, CAD-SAM, and Simple K-means

S&P 500

Dataset	Algorithm	Recall (%)	Precision (%)	F2 (%)	F4 (%)
Consumer Staples Weekly	CAD	34.7	0.33	1.59	4.86
	CAD - SMA	95.05	4.7	19.62	44.61
	Simple K-means	40.01	3.8	13.77	25.64
Consumer Discretionary Weekly	CAD	34.15	0.33	1.6	4.88
	CAD - SMA	88.03	1.5	7.02	20.04
	Simple K-means	40.7	3.6	13.30	25.34
Energy Weekly	CAD	34.49	0.33	1.58	4.83
	CAD - SMA	86.7	7.4	27.58	53.18
	Simple K-means	40.2	1.5	6.53	15.97
Financial Weekly	CAD	35.47	0.34	1.65	5.05
	CAD - SMA	97.23	2.9	12.95	33.37
	Simple K-means	40.74	3.62	13.35	25.41
IT Weekly	CAD	33.69	0.34	1.63	4.98
	CAD - SMA	96.41	6.92	26.88	54.76
	Simple K-means	18.32	0.83	3.51	8.18
Consumer Staples Daily	CAD	32.11	0.3	1.43	4.39
	CAD - SMA	96.52	1.03	4.94	14.96
	Simple K-means	38.46	1.52	6.56	15.83
Consumer Discretionary Daily	CAD	34.91	0.34	1.65	5.03
	CAD - SMA	93.88	1.03	4.93	14.90
	Simple K-means	42.85	1.5	6.58	16.35
Energy Daily	CAD	32.42	0.32	1.54	4.7
	CAD - SMA	94.61	0.91	4.38	13.41
	Simple K-means	50	1.49	6.66	17.15
Financial Daily	CAD	33.96	0.31	1.5	4.61

	CAD - SMA	99.57	0.99	4.76	14.52
	Simple K-means	42.85	0.74	3.06	6.88
IT Daily	CAD	32.58	0.31	1.5	4.6
	CAD - SMA	96.42	1.06	5.08	15.32
	Simple K-means	14.28	1.32	5.88	15.03

Table A.2. Complete Results of CAD, CAD-SMA and Simple K-means Applied to QSM Sectors

Qatar Stock Exchange					
Dataset	Algorithm	Recall (%)	Precision (%)	F2 (%)	F4 (%)
Financial Weekly	CAD	26.4	7.3	17.33	7.56
	CAD-SMA	92.8	0.57	2.78	8.82
	Simple K-means	40	1.08	4.87	1.13
Consumer Goods & Services Weekly	CAD	28	7.68	18.31	7.95
	CAD-SMA	98.01	0.5	2.45	7.86
	Simple K-means	33.33	0.42	2.00	0.44
Industrial Weekly	CAD	28	4.83	14.29	5.03
	CAD-SMA	98.67	0.71	3.45	10.82
	Simple K-means	33.33	0.44	2.09	0.46
Insurance Weekly	CAD	32.67	4.96	15.43	5.17
	CAD-SMA	98.6	0.8	3.87	12.04
	Simple K-means	33.33	0.59	2.75	0.62
Financial Daily	CAD	16.49	1.84	6.36	1.92
	CAD-SMA	96	1.21	5.76	17.12
	Simple K-means	18.18	0.42	1.92	0.44
Consumer Goods & Services Daily	CAD	18.86	2.79	8.76	2.91
	CAD-SMA	94	1.12	5.35	15.99
	Simple K-means	28.57	0.51	2.38	0.53
Industrial Daily	CAD	20.13	1.17	4.75	1.22
	CAD-SMA	94.67	1.02	4.89	14.79
	Simple K-means	25	0.23	1.11	0.24
Insurance Daily	CAD	32	2.12	8.38	2.22
	CAD-SMA	99	1.1	5.27	15.88
	Simple K-means	16.67	0.33	1.53	0.35

APPENDIX B: DATASET SAMPLES

Dataset B.1. Sample of Weekly Stock Prices of S&P Information Technology Sector

Name	APPLE	MICROSOFT	INTERNATIONAL BUS.MCHS.	ORACLE	GOOGLE 'A'
Code	@AAPL	@MSFT	U:IBM	U:ORCL	@GOOGL
5/6/14	594.4099	39.06	190.03	41.01	522.5698
4/29/14	592.3298	40.51	195.11	40.11	536.3298
4/22/14	531.699	39.99	192.15	40.46	545.5
4/15/14	517.9597	39.75	197.02	39.73	548.7
4/8/14	523.4399	39.82	193.29	40.24	557.5098
4/1/14	541.6499	41.42	194.5	41.49	567.9954
3/25/14	544.99	40.34	195.04	38.4	579.9219
3/18/14	531.3999	39.55	186.81	38.84	606.2175
3/11/14	536.0898	38.02	186.76	38.9	600.5769
3/4/14	531.24	38.41	186.44	39.41	608.0442
2/25/14	522.0598	37.54	183.23	38.25	610.5918
2/18/14	545.99	37.42	183.19	37.97	606.0273
2/11/14	535.96	37.175	179.7	37.84	595.6672
2/4/14	508.7898	36.35	172.84	35.96	569.6321

Dataset B.2. Sample of Weekly Stock Prices of S&P Information Technology Sector with Change Percentage Preprocessing

Name	APPLE	MICROSOFT	INTERNATIONAL BUS.MCHS.	ORACLE	GOOGLE 'A'
Code	@AAPL	@MSFT	U:IBM	U:ORCL	@GOOGL
5/6/14	0%	-4%	-3%	2%	-3%
4/29/14	11%	1%	2%	-1%	-2%
4/22/14	3%	1%	-2%	2%	-1%
4/15/14	-1%	0%	2%	-1%	-2%
4/8/14	-3%	-4%	-1%	-3%	-2%
4/1/14	-1%	3%	0%	8%	-2%
3/25/14	3%	2%	4%	-1%	-4%
3/18/14	-1%	4%	0%	0%	1%
3/11/14	1%	-1%	0%	-1%	-1%
3/4/14	2%	2%	2%	3%	0%
2/25/14	-4%	0%	0%	1%	1%
2/18/14	2%	1%	2%	0%	2%
2/11/14	5%	2%	4%	5%	5%

Dataset B.3. Sample of Weekly Stock Prices of S&P Information Technology Sector with SMA Preprocessing

Name	APPLE	MICROSOFT	INTERNATIONAL BUS.MCHS.	ORACLE	GOOGLE 'A'
Code	@AAPL	@MSFT	U:IBM	U:ORCL	@GOOGL
5/6/14	1.2%	0.5%	0.7%	1.0%	-0.6%
4/29/14	1.3%	0.9%	1.0%	0.9%	-0.4%
4/22/14	0.4%	0.8%	0.9%	1.0%	-0.4%
4/15/14	0.2%	0.9%	1.3%	1.0%	-0.3%
4/8/14	0.3%	1.0%	1.2%	1.2%	-0.2%
4/1/14	0.7%	1.5%	1.4%	1.7%	0.0%
3/25/14	0.9%	1.4%	1.6%	0.8%	0.2%
3/18/14	0.6%	1.3%	1.2%	1.1%	0.9%
3/11/14	0.9%	0.8%	1.3%	1.4%	0.9%
3/4/14	0.9%	1.1%	1.6%	1.9%	1.3%
2/25/14	0.7%	0.8%	1.5%	1.6%	1.8%
2/18/14	2.4%	1.0%	2.0%	1.9%	2.1%
2/11/14	2.7%	1.1%	2.0%	2.6%	2.3%

APPENDIX C: CODE SAMPLES

Code C.1. Anomaly Detection Code Sample (Using R)

```
predictAnomalies <- function(){

  start = overlap
  end = nrow(myData)
  centroid = numeric(end)

  windowStart = overlap
  windowEnd = ncol(myData)

  # calculate centroid of companies in sector
  for(i in 1:end){
    centroid[i]=mean(as.numeric(myData[i,]))
  }

  while( windowStart < windowEnd){

    windowStart = windowStart - overlap + 1
    windowStop = windowStart + windowSize - 1

    if(windowStop > windowEnd)
      windowStop = windowEnd

    for(i in 1:ncol(myData)){

      # get pearson correlation of company i and centroid
      ci = cor(centroid, myData[,i])

      for(j in 1:nrow(myData)){

        predictedX = myData[j-1,i]*ci

        error = sqrt((predictedX-myData[j,i])^2)

        # if errorgreater than standard deviation then it is an anomaly
        if(error > stdev[i]){
          predictLabel[j,i] <<- "o"
          predRow[j] <<- "o"
        }
      }
    }
    windowStart = windowStop
  }#end while
}#end function
```

Code C.2. Anomaly Insertion Code Sample (Using R)

```
insertAnomalies <- function(){

  per = 0.1 #percentage of Anomalies to be inserted

  #for every column (company) find statistical variables
  for(i in 1:len){
    skew[i] = skewness(myData[,i])
    stdev[i]=sd(myData[,i])
    avg[i] = mean(myData[,i])

    if(skew[i] == 0){
      upperbound[i] = avg[i] + (3*stdev[i])
      lowerbound[i] = avg[i] - (3*stdev[i])
    }else{
      Q1[i] = quantile(myData[,i],.25)
      Q3[i] = quantile(myData[,i],.75)
      IQR[i] = IQR(myData[,i])

      upperbound[i] = avg[i] + (Q3[i] + (3*IQR[i]))
      lowerbound[i] = avg[i] - (Q1[i] - (3*IQR[i]))
    }
  }# end for

  per = per/100

  numAnom = as.integer(per*(nrow(myData)*ncol(myData)))

  for (i in 1:numAnom){

    randomRow = sample(1:nrow(myData), 1)
    randomCol = sample(1:ncol(myData), 1)

    upOrLow = sample(1:2, 1)

    if (upOrLow == 1) {
      #print("anomaly is set to upperbound of value")
      myData[randomRow,randomCol] <<- upperbound[randomCol]
    }else{
      #print("anomaly is set to lowerbound of value")
      myData[randomRow,randomCol] <<- lowerbound[randomCol]
    }
    trueLabel[randomRow,randomCol] <<- "o"
    labelRow[randomRow] <<- "o"
  }# end for
}# end function
```