Financial Market Service Architectures: A “Pump and Dump” Case Study

Mohamed Zaki, David Diaz, and Babis Theodoulidis
Centre for Service Research
Manchester Business School
University of Manchester
Manchester
United Kingdom
mohamed.zaki@postgrad.mbs.ac.uk, david.diaz@postgrad.mbs.ac.uk
b.theodoulidis@mbs.ac.uk

Abstract— This paper describes a service architecture for a financial market monitoring and surveillance system in which different components interact in coordination with internal and external service providers to produce proactive alarms for potential fraud cases. The proposed service system is demonstrated through an exemplar case study of text mining and data mining to analyze the impact of ‘stock-touting’ spam e-mails and misleading press releases on trading data. It also shows how an independent service provider could have helped by raising the alarm about a potential on-going ‘pump and dump’ scheme. The proposed service architecture extends the Market Monitoring Framework (MMF) [6], by incorporating automated linguistics-based text mining techniques to extract the key concepts of spam e-mails and press releases, relate them with other available information and highlight any signs that they might be part of ‘stock touting campaigns’. The analysis of emails and press releases through text mining components could help to raise proactive alarms that would not have been possible otherwise. Evaluation of the proposed service system is carried out through a case study that relates to a real case from the over-the-counter (OTC) market, and which was prosecuted by the SEC. Through this case, the paper goes on to explain how the proposed approach could be used within the existing fraud analysis process, the extent to which the process could be automated, the relationships with other types of analysis and the role that fraud analysts play.

Keywords- Service systems, Spam E-mails, Pump and Dump manipulation, Fraud Detection, Text Mining, Financial Market Monitoring system

I. INTRODUCTION

Microcap fraud is an on-going market manipulation scheme that presents several challenges for investors and financial authorities alike. Toughened rules and regulations have recently been put in place and authorities have taken drastic action against fraudsters. The existing literature shows that most manipulation cases occur in unregulated markets such as OTC and Pink Sheets because they are relatively small, illiquid, involve fewer disclosure requirements for listed firms, are prone to conflicts of interest, lack accurate public information [8], and are subject to less regulatory oversight [1].

In this sense, the continuous improvement and development of financial market surveillance is essential to guaranteeing an efficient market. In a recent consultation, the European Commission and the US Securities Administrators recommended that regulators pursue further research into the development of a comprehensive surveillance system to supplement current market surveillance. This paper extends the Market Monitoring Framework (MMF) [6] by incorporating automated, linguistics-based text mining analysis to help raise proactive alarms and highlight potentially fraudulent patterns generated as part of a ‘touting campaign’. Additional components have been added to the MMF framework to address requirements for proactively identifying possible fraudulent “pump and dump” manipulation scheme. The paper utilizes a service system approach where the internal and external service providers interact to produce proactive alarms. In this context, it would have been difficult to detect this manipulative scheme without these components. In fact, given the strategy followed by the manipulators, it was not possible to raise early alarms of manipulation, and the fraud related actions were detected only when the actual manipulative trading occurred.

Indeed, accurate information is the best tool to enable wise investment in the financial market. However, due to its scarcity, investors (and microcap investors in particular), tend to rely heavily on unstructured sources of information such as unconfirmed press releases, forums, and spam e-mails among others. More specifically, fraudsters pump microcap stocks by releasing false messages that urge investors to buy or sell particular stocks. The fraudsters then make significant profits by dumping the stock when the prices rise artificially. This is especially true for unregulated markets, such as OTC, which are extremely illiquid, and where investors face the additional risk of investing in smaller or startup companies which are in the majority in this type of market. Thus, these unstructured sources could be used to create proactive indicators and alerts of potential microcap manipulation cases, especially if these events are accompanied by changes in trading behaviour. Using on-demand analysis services could help investors and fraud analysts to raise alarms earlier. The novelty of this work is to demonstrate the proposed approach and explain how it could be used within the existing fraud analysis processes, the extent to which the processes could be automated, the
relationships with the other types of analysis and the role played by fraud analysts.

A further contribution of this paper is to identify the gaps in financial research studies. None of the existing works has considered the use of automated, linguistic-based text mining techniques for the analysis of ‘stock touting’ in unregulated markets. In particular, the paper demonstrates the impact of using text mining techniques to automate analysis and classification of large amounts of financial textual sources, identify interesting patterns and relationships between them and help fraud analysts to proactively investigate possible cases of manipulation. This could help regulators and market participants to minimize the risks from ‘pump and dump’ schemes.

The paper is organized as follows: the second section presents a brief literature review on the relevant topics. The third section discusses the methodology and instantiation of a case study on the market monitoring framework. The fourth section presents and discusses the results. Finally, the fifth section summarizes this paper and outlines directions for future work.

II. EXISTING WORK
There are two broad research areas of relevance to this work. The first focuses on spam, as numerous pieces of research have identified spam e-mails as a central problem that requires further investigation. Many authors have reported on the presence of abnormal returns, higher volumes or episodes of unusual volatility as a consequence of stock touting campaigns, with Aggarwal and Wu’s study among the most influential [1].

A number of studies have been carried out on touting campaigns which use the internet, and particular emphasis has been placed on the effects of stock touting spam e-mails. For example, Mei, Wu, and Zhou [13] present some empirical evidence of an equilibrium model to demonstrate how “smart money” can profit from other investors’ irrational behaviors. The model is based on a sample collected from cases which have been prosecuted by the SEC. The suggested model classified investors into behavior-driven investors, arbitrageurs, and a manipulator. The manipulator’s strategic action, together with other investors’ behavioral biases can lead to higher returns, increased volatility and trading volumes, short-term price contribution and long term price reversal. In particular, the effect of these rumors and touting messages is more evident in the unregulated market that endogenously drives pricing of several securities up and affects market integrity [13].

A further study by Böhme and Holz found evidence of the harmful effect of spam messages on the financial markets [2]. The study employed a multiplicative multivariate regression model and a classical event study methodology, focusing on the effect of spam e-mails on the return and volume of the target stocks.

The work of Frieder and Zittrain [9] evaluated and analysed the impact of touted spam on the trading activity of specific stocks using a broad sample from the Pink Sheets market. The paper compared the touted stock with another control sample (not touted) during the same period. Evidence was found of a significant positive return on days where heavy spam touting took place. Furthermore, the volume of trading is shown to correspond positively to heavy spam touting.

Hanke and Hauser’s analysis [11] begins with a description of the common characteristics of the advertised stock, such as price level and average turnover. Besides investigating the effect of touting spam on returns and volume, they also measured the effect of spam email on other variables such as excess returns, turnover and intra-day volatility. The research showed that stock touting had a significant positive impact on securities prices. Furthermore, liquidity is the one of the foremost factors in determining the success of spamming campaigns. Lastly, repetitive spam sent on consecutive days continued to increase demand for the targeted stock, which strengthened the spammer’s position and afforded them more time for liquidation.

A particularly interesting piece of research studied the 'attention effect' in the market response to spam messages. The research showed that returns and abnormal volume were highest when spam e-mails cited stocks with extreme target prices and copied press release information. Furthermore, the 'attention role' helped spammers select the target stocks. The study emphasized that successful touted email campaigns have an influential impact on investor attention and affect the asset prices. Regression analyses showed that messages specifying a target price trigger a significant market response when accompanied by information from a press release [14].

Therefore, the dramatic growth of the Internet has transformed the way investors receive information about publicly traded companies. As is evident from the earlier literature, various studies provide anecdotal evidence that touting campaigns using spam e-mails engender substantial fluctuations in price, volatility, returns and trading volume during stock campaigns, and that there are some recipients who read the messages and act upon them. However, the analysis and evaluation performed in previous research followed a traditional statistical approach and relied heavily on manual annotation and classifications.

As discussed above, the existing literature has not considered the use of text mining techniques for the analysis of “stock touting” in unregulated markets. Given that the quantity of spam e-mails is increasing exponentially, there is a clear need for an automated linguistic-based text mining system to extract key concepts from these textual sources, and link it with structured data to provide better analysis and an accurate securities fraud detection model. In particular, the SEC is concerned about the growth of these sources of Internet fraud because they could represent successful attempts to manipulate stock prices. Therefore, this paper used the MMF to analyze these information-based manipulation sources and demonstrate how automated linguistic-based text mining helps to raise proactive alarms about touting stocks and make fraud analysts aware of potential manipulative practices. The text mining employed linguistic techniques and incorporated domain specific resources such as dictionaries, taxonomies and financial ontology to assist in the extraction analysis.
Extensive research on the use of text mining algorithms to filter spam e-mails has been reviewed by [19]. [19] discussed the methods have been used by previous research; the first method is blocking the sources of spam based on features such as e-mail address and e-mail servers. The second method is analysing the content of spam e-mails by employing the rule-based method. Content-based filtering was another approach used to classify models and generate automatic content filtering rules.

The service system used the proposed systematic framework of market monitoring (MMF) and considered the current and potential characteristics of market monitoring systems to demonstrate how text mining analysis integrates with existing data mining techniques to generate a better detection model and help prosecute cases of fraud. The system incorporates data mining techniques used by several authors, such as Goldberg, Kirkland et al [10] and Kirkland and Senator [12], who described fraud detection systems used to monitor trades and quotations in various financial markets such as NASDAQ. In particular, techniques such as decision trees, neural networks and linear regression are considered to analyze prosecuted cases in regulated markets [5] and unregulated markets [19].

The following section demonstrates the service system proposed to investigate a “pump and dump” case study from the OTC Pink Sheets market.

III. METHODOLOGY

In this paper, the different tasks and analysis components were developed following the guidelines of the CRISP Data Mining reference model [3] using the IBM-SPSS PASW13 data and text mining workbench.

A. Business Understanding

This section discusses the analyses of the case study shown in Figure 1 to demonstrate the fraud case. In this particular case, fraudsters created artificial demand on Mobile Ready Entertainment Corp. stock (SYMBOL: MRDY), a publicly-traded company quoted on the Pink OTC Markets, Inc. (‘Pink Sheets’). This was achieved through the distribution of false and misleading press releases, and possibly through the distribution of spam e-mails. The investigation found that several press releases contained disingenuous revenue projections, and mentioned indistinct business relationships. As depicted in Figure 1, fraudsters issued two misleading press releases in January and February 2007 respectively. On March 8, 2007 and the following days, an unidentified agent sent several spam e-mail messages. Furthermore, on March 10, 2007, an independent stock spam tracker, called SPAMNATION.info, reported on the presence of such a touting campaign. On an undetermined later date, Craig A. Mora, the president of the company, sent a disclaimer e-mail to SPAMNATION.info stating that ‘... spam promoting MRDY.PK: was sent out by an investor trying to create a market to sell to. The company was in no way responsible for nor connected to the spam mail’.

After the spam e-mail campaign, the record shows that a number of copies of Form 144 were completed and submitted to the SEC in April and June 2007, indicating the intention of the company’s directors to sell their shares. The forms were registered after the shares had actually been sold. Following the completion of the Form 144s, a number of misleading press releases were issued on consecutive days towards the middle of July 2007. These press releases contained false statements of material fact and mentioned a contract with a company called ‘Simply Fit Inc.’

The findings detailed in this paper correspond with what was reported by the SEC investigation. In July 2008, the SEC published a case against the fraudsters which stated that between January and July 2007, the defendants directly and indirectly made use of sixteen separate press releases, many of which contained false statements of material fact. For example, on July 19, 2007, manipulators issued a press release falsely claiming that Mobile Ready had been ‘awarded an exclusive contract with Simply Fit Holdings, Inc.’ Furthermore, they claimed that the contract with Simply Fit Holdings Corporation could yield in excess of one million ($1,000,000) dollars of revenue for [Mobile Ready] over the next year...’ The two managers of the company had personally received 76.9 million Mobile Ready shares. However, these shares had never been registered with the Commission or subject to an applicable exemption from registration. While the market for Mobile Ready shares was artificially inflated, they began to sell shares into the public market [15]. However, the SEC case does not mention the role played by spam e-mails (if any), in the touting campaign.
“Between April and July 2007, Magolnick sold 2,100,333 shares, Mora sold 2,249,833 shares of Mobile Ready” (Sec 2008)

Figure 1. MRDY Case Timeline

Regulating Authority, Market Monitor, Trader, Public, Fraud Analysts.

Figure 2 Fraud Detection Service System (MMF Instantiation)
B. Data Sources

The case study uses various sources such as real trading data retrieved from the Bloomberg data service provider, as well as a collection of unstructured data sources related to the MRDY Company. For structured data, the research had access to MRDY closing prices, including bids and asks information, for the whole of 2007. In addition, the research uses a spam e-mail dataset collected from Richardson’s Stock Spam Effectiveness Monitor (SSEM) archive [4]. This source contains spam messages from the period January 2006 to February 2008. Figure 3 shows an example of an MRDY spam e-mail found in the database. This study also used the spam watcher list available at spamnation.info, containing a list of companies whose stocks where touted during the year 2007.

Figure 3 An example of an MRDY Spam E-mail

The study also includes information relating to news and financial events collected from the Factiva database [7]. This included press releases, new products/services, corporate industrial news, branding, corporate digests and other types of news posted in relation to the touted company. Figure 4 presents an example of an annotated press release that was issued on July 19, 2007.

Figure 4 An example of a MRDY Press Release

The Thomson Reuters Insider Filings Data Feed (IFDF) is used as a source for all U.S. insider activity as reported on Forms 3, 4, 5, and 144. The research collected a total of 37 items, all of them corresponding to Form 144 filings. The IFDF does not contain any other insider filing forms about Mobile Ready Co. for 2007. The IFDF is available from the Wharton Research Data Services [16]. Finally, the EDGAR database on the SEC website was used as a source of information regarding the proceedings in actions relating to the MRDY case. The information can be found at http://www.sec.gov/litigation/complaints/2008/comp20644.pdf. The MRDY case was selected for this research because it presents an instance of a ‘pump and dump’ microcap manipulation which matches the study interest as it highlights the need for automated, linguistics-based text mining. In addition, the case demonstrates how the fraud detection services prototype could play a vital role in proactively monitoring, detecting, and deterring market abuses in the OTC markets.

C. Data Preparation

Fraud detection systems can contain a retroactive or proactive monitoring engine that utilizes different types of data source [6]. Data could initially be explored by an Information Management component where different processes for data preparation take place. Restructured data, along with updated patterns and models could be used in this layer to perform a wide range of analysis, which can run either automatically or on an on-demand basis. More specifically, as shown in Figure 2, the Information Management layer includes a range of data preparation components including normal database manipulation operations components, such as query and OLAP reporting, as well as other data and text mining components.

Before any analysis takes place, the data need to be pre-processed in order to combine and match different sources. These can either be structured, such as trading data, or unstructured, such as spam e-mails, spam tracker feeds, press releases and financial filings. However, as the pre-processing and extraction of key terms is a difficult task, the analyst could use externally provided analysis to cross-check their own internal results. For instance, spam e-mails are a very dirty, unstructured source of data and the instantiation incorporates a dedicated set of tasks to improve the quality and address any inconsistencies, inaccuracies and omissions in the original data. Furthermore, the instantiation considers the option of including input data from external unstructured data analysis providers: one dedicated exclusively to the analysis of spam e-mails and e-mail traffic, and another to the analysis of news and press releases which come from formal sources.

The following section demonstrates the deployment of fraud detection system and the instantiation of a market monitoring system. In particular, the objective of this instantiation is to demonstrate how existing detection systems which use data mining can be enhanced. The analysis shows how an automated, linguistics-based text mining approach could help to proactively raise alarms for possible ‘pump and dump’ manipulation through the use of unstructured data sources. Using on-demand analysis services could help investors and fraud analysts to raise alarms earlier and prevent potential manipulations.
IV. **MMF Instantiation as a Fraud Detection Service System (Deployment)**

This section demonstrates the deployment and the instantiation of a market monitoring system from the perspective of an independent provider of fraud detection services. Indeed, the work considers the use of different sources of financial and textual data to discover patterns in trading indicators related to market manipulation activities. The independently provided fraud detection system [6], shown in Figure 2, combines various techniques of text mining and data mining to analyze structured and unstructured data. These techniques can be performed either inside or outside the main system boundaries. Furthermore, linguistic resources components such as financial ontology, market manipulation, taxonomy, and financial dictionaries have been used to semantically enrich the knowledge base in the MMF. In particular, these linguistic resources could play a vital role between information management and business intelligence analysis components.

In this instantiation, structured data preprocessing and outlier detection tasks are performed inside the system. In particular, trading data is used to detect unusual trading volumes, jumps in prices and imbalances between bids and asks in trading orders. Unstructured data is also processed internally, with the option of requesting additional external services analysis. In this sense, the analyst has the choice of integrating information provided by external service agents or relying solely on the system’s internal capabilities. Unstructured sources, such as spam-mails and filing forms, could be used to create proactive indicators, alerts and attention-grabbing devices for potential microcap manipulation cases, especially if these events are accompanied by significant changes in trading behavior.

A. **Information Management**

As shown in Figure 2, the Information Management layer contains OLAP, outlier data mining detection, association analysis and visualization components. If we turn our attention to the timeline of the case, in Figure 5 we can appreciate that the outlier detection component, based on the proposed financial indicators introduced by Zaki, Theodoulidis and Díaz [19], raised alarms in close relation with jumps in volume while the trading of the insiders was executed. For example, it is possible to distinguish alarms, represented as red squares on Figure 5, on and after July 24, 2007. However, it is important to point out that the touting campaign started early in January 2007 with the release of misleading press releases and, arguably, spam e-mails. Thus, the instantiation shows the importance of using on-demand analysis of unstructured data that could provide proactive alerts to analysts and prevent manipulation.

Figure 5 Volume and Outlier Detection Alarms

Figure 6 shows some attributes of Form 144 which include ticker, insider name (owner), the date of filing (fdate), the date SEC received the form (secdate), the expected date of sale (psaledate), date of insider signature on the form (sigdate), the number of securities to be sold (pshares), the estimated market value of the proposed sales (value), and the name of the executing broker (broker).

<table>
<thead>
<tr>
<th>broker</th>
<th>ticker</th>
<th>fdate</th>
<th>secdate</th>
<th>psaledate</th>
<th>sigdate</th>
<th>pshares</th>
<th>value</th>
<th>broker</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROBY BAGOLINICK MICHAEL</td>
<td>2007-04-11</td>
<td>2007-06-02</td>
<td>2007-04-01</td>
<td>221850X</td>
<td>448500</td>
<td>0.000</td>
<td>221850X</td>
<td>448500</td>
</tr>
<tr>
<td>ROBY BAGOLINICK MICHAEL</td>
<td>2007-04-08</td>
<td>2007-06-27</td>
<td>2007-06-19</td>
<td>123790X</td>
<td>474920</td>
<td>0.000</td>
<td>221850X</td>
<td>448500</td>
</tr>
<tr>
<td>ROBY BAGOLINICK MICHAEL</td>
<td>2007-07-06</td>
<td>2007-07-09</td>
<td>2007-06-31</td>
<td>950293</td>
<td>251600</td>
<td>0.000</td>
<td>221850X</td>
<td>448500</td>
</tr>
<tr>
<td>ROBY BAGOLINICK MICHAEL</td>
<td>2007-08-17</td>
<td>2007-08-08</td>
<td>2007-08-15</td>
<td>2208350</td>
<td>596820</td>
<td>0.000</td>
<td>221850X</td>
<td>448500</td>
</tr>
<tr>
<td>ROBY BAGOLINICK MICHAEL</td>
<td>2007-09-05</td>
<td>2007-09-05</td>
<td>2007-09-05</td>
<td>950293</td>
<td>251600</td>
<td>0.000</td>
<td>221850X</td>
<td>448500</td>
</tr>
<tr>
<td>ROBY BAGOLINICK MICHAEL</td>
<td>2007-11-19</td>
<td>2007-11-08</td>
<td>2007-11-15</td>
<td>378080X</td>
<td>338300</td>
<td>0.000</td>
<td>221850X</td>
<td>448500</td>
</tr>
</tbody>
</table>

Although the fraudsters reported their intention to sell, as stated on the SEC proceeding and corroborated by the Form 144 information, submission of the form occurred after the actual sale of the shares, as shown in Figure 6. Moreover, looking at Figure 7, we can infer that an association analysis performed using the Form 144 information, does not reveal any strong relationships between dates, insiders and brokers. Thus, analysis of the filing form alone would not have been sufficient to detect the on-going fraud scheme.

B. **Text Mining Analysis**

As mentioned before, unstructured data could be processed internally, externally or internally with cross validation from external service providers. In Figure 2, the

Figure 6 OLAP analysis IFDF (Form 144)

Figure 7 Association Analysis of IFDF (Form 144)
unstructured data analysis is depicted as an internal process with the option of cross validation of the external data service providers. Furthermore, the instantiation shows the importance of using on-demand text mining analysis to analyze unstructured data that could provide proactive alerts in the output management. This could help fraud analysts to investigate and prevent manipulation. The fraud analysts could keep the system watch list updated by touted stocks in the “Knowledge Update”.

Stock spam e-mail is a good example of very unclean, unstructured data. Thus, the proposed text mining methodology employs two methods from the information extraction process: namely, dictionary-based and pattern-based. The dictionary-based information extraction method involves configuring and training systems with linguistic resources such as thesaurus, list of terms, concepts, synonyms, and types (semantic groupings of concepts). This method could use tokenisation and morphological analysis to extract terms from spam e-mails and match them with the appropriate named entities. Tokenisation analysis parses documents into characters and words that are called tokens. Text mining systems make use of advanced algorithms that deal with tokenisation challenges, such as inconsistent punctuation, and special characters such as dash symbols and apostrophes among others. For the morphological analysis, stemming algorithms are employed to identify the root of terms listed in the dictionary. Text mining applications integrate stemming algorithms on the tokenisation output to conflate the tokens into an orthogonal set of distinct morphological groups that are used to train the extraction engine to group similar forms (singular or plural) of terms and add them to the dictionary. However, this method is time consuming and can produce inaccuracies if the terms are not appropriate to the domain of financial analysis.

The prototype developed a library comprising an updated list (October 2009) of 23,028 over the counter (OTC) traded stocks (Symbol ticker, company name). In addition, the Pink Sheets market supplied a list of symbol ticker and company names issued between 2006 and October 2009. Regarding the market data, the data preparation phase involved appending and sorting activities to construct the final dataset, i.e., data from the initial raw data that was fed into the analysis. The OTC and Pink Sheets data sources were appended to create one list of issued stock symbol tickers and a company list to be used in the analysis phase. Restructured market data and converted e-mail documents are stored in a knowledge base to feed the taxonomy analysis with a clean dataset. For example, the symbol ticker is made up of four or five characters. During the extraction process, the e-mails were scanned and analysed in order to identify symbol tickers within the dictionary, and map the symbols to an event called <Stock Symbol>.

However, spammers could deceive the automatic filtration systems through tactics such as changing the way they cite the symbol e.g. (MRDY.PK, M-R-D-Y, M R D Y, etc) which obviously could not be identified using the dictionary method. Therefore, pattern-based information extraction could help to address these challenges by employing natural language processing (NLP). This method extracts information from e-mails using linguistic patterns or rules to recognise features of interest, especially unrecognised concepts from the dictionary-based model. This method performs a deeper analysis on the words, phrases and syntax inside the e-mails, and thus helps to uncover knowledge of the underlying language in these messages.

These patterns are designed as a set of rules that describe structures in words, combinations of structures or combinations of rules. For example, if the task is to find or recognise stock symbol tickers that have been mentioned in the e-mail body or text, it is possible to define a rule in which all collections of four to five capital characters are mapped to an entity called <Stock Symbol>. An advantage of this approach is that it is possible to extract not only stock tickers that were listed in the dictionary originally, but also to extract newly issued symbol tickers that match the predefined rules. Figure 8 shows how effective the pattern-based method was in extracting stock symbols from spam e-mails in 2007.

![Figure 8 Stock Symbol Information Extraction Comparison](image)

Furthermore, the pattern-based method incorporates a named entity recognition analysis that identifies and extracts certain non-linguistic terms and maps them to entities such as date, person, e-mail address (sender and receiver), and IP address which will be used to extract the metadata associated with spam e-mails. The subject line of the e-mail is another interesting element for analysis, as it could be one of the features which prompt investors to open and read the e-mail. Overall, while it is very useful to analyse the metadata from spam e-mails, it could contain misleading or incorrect information. For example, date stamp analysis can be performed to determine when spammers started their campaign for a specific stock, and how long they continued their campaign for that stock, for example.

Moreover, the pattern-based method utilises word association analysis to identify the relationship between two or more entities. For example, through manual annotation we have observed that the word ‘symbol’ is generally followed by a symbol ticker, for example Symbol:MRDY, as shown in Figure 9. Additionally, different forms of the word symbol have been taken into account, such as, ‘stock symbol’, ‘target symbol’, ‘symbol’, ‘sym’, ‘symsymbol’, ‘o.t.c. sym bol’, ‘o.t.c symbol’, among others.
The analysis generates the required stock spam e-mail taxonomy to extract the key concepts and attributes required for fraud detection as shown in Figure 10. The metadata analysis deals with key header information such as date, time, sender, receiver, and subject. The body analysis deals with extracting named entities of stock spam e-mails for further analysis. For example, it extracts information on stock profiles such as the symbol ticker, the company holding the stock, and the sector that the company is involved in, as well as stock quote information that could tempt investors, such as target price or volume projections, trading date expectations, recommendations, and financial indicators such as buying or selling signals. Furthermore, the application aims to extract the concepts and phrases that indicate whether or not the cited news relates to the stock.

Previous papers by Nelson, Price and Roundtree [14], argued that target price quotes and copied press release information advertised in spam e-mails could trigger a significant market response. Therefore, this research extracted target stock price and volume, financial long or short term investment indicators, speculative trade date, and buy or sell signals contained in the spam e-mails. In addition to the body content extraction, metadata from the spam e-mails was extracted and analysed. In particular, date stamp analysis was used to determine when spammers started their campaign for MRDY, and how long they continued their campaigns for.

The result of this process is a table containing the symbol, metadata information (date, sender), and body copy information such as current price, target prices and speculative potential profit, as shown in Figure 11. The MRDY spam e-mail campaign was distributed on March 8, 9 and 10, and April 7. The analysis shows that MRDY was heavily touted on March 9, (24% of the total spam messages) and on March 10, 2007 (72%). Furthermore, spammers recommended and encouraged investors to buy and hold the stock, and referred investors to the misleading press release that had been issued previously. For example, spammers used push phrases such as ‘check or see the news’, ‘the hottest news released’, ‘see bullish news online’, and ‘fresh news’.

In order to provide cross validation the analysis incorporated information from the SPAMATION website. This website provides feeds that include information about stocks advertised in unsolicited commercial e-mails. In 2007, MRDY was listed as a touted stock on their database, as shown in Figure 12. However, the site does not provide detailed analysis of the content or metadata of spam e-mails.

Press releases and news issued in relation to the Mobile Ready company in 2007 have been retrieved. News items maintained their original XML tags and XML analysers were used to query key concepts and metadata provided in these tags such as symbols, company name, date, subject, and other information cited in the body. Based on the analysis of news, Figure 13 shows press releases were heavily distributed in July 2007, especially on the 16th, 17th, 18th, 19th and 25th. In line with the SEC case, the defendants
were trying to sell their shares. Therefore, they issued and posted press releases to mislead investors and pump the MRDY price up, then dump it again after they had sold their shares. For example, as shown in Figure 4, they released misleading information regarding the impact of the contract with Simply Fit. Sentences like: 'Simply Fit gears up to capture a share of $100 billion in annual beverage industry sales' or 'The project could yield in excess of one million ($1,000,000) dollars of revenue for MRDY over the next year' could be found in the news text.

Regarding the press release analysis, in this instantiation we did not implement sophisticated text mining extraction analysis, and we only considered the distribution of the releases. Automatic analysis and classification of news items and press releases will be implemented in future research.

C. Economic Analysis

Our economic analysis used the output from the Information Manager layer to create a complete representation of the fraud scheme. Figure 14 shows a graphical representation of the events using the tools available from the pattern visualisation analysis component. The visualisation includes the price and volume time series, plotted together with the outliers' alarms, cumulative news, and cumulative volume sales as declared on the Form 144. From left to right, we can see that alarms were always reactive to the changes in volume, and occurred several days after the misleading press releases were issued. The only proactive signals available throughout the duration of the manipulative scheme came from the body content analysis of spam e-mails which were sent at the beginning of March (see Figure 8), at least two months after the first misleading press releases. In fact, there was a slight increase in the stock price associated with the spam e-mail campaign, reaching a high of $0.06 on March 12 and 13, 2007. However, this increase in price was not accompanied by a significant increase in volume. According to financial theory, markets respond to new information, and as expected, there is some discernable reaction or effect associated with the press releases. Thus, the press releases sent on July 16, 17, 18, 19, and 25 led to an increase in the volume of trading, which was also captured by the outlier detection engine, raising corresponding alarms. In summary, considering all the information available, we can stress that it would have been difficult to detect this manipulation scheme without the spam analysers, and potentially, the press release analyser components. This type of unstructured source is also more difficult to analyse, and thus, on-demand analysis services from external providers could have helped fraud analysts raise alarms about suspiscious behaviour more than one year before the actual SEC case prosecution was completed.

V. CONCLUSIONS

This paper proposes the extension of the Market Monitoring Framework (MMF) with a number of additional components that address requirements for proactively identifying possible fraudulent “pump and dump” scenarios through the analysis of textual information resources. The paper utilizes a service system approach where the internal and external service providers interact to produce proactive alarms. The paper demonstrates the use of the prototype detection system through a case study investigation of a real case that took place on the Pink Sheets market during the first half of 2007. The research demonstrates that the use of additional sources of information such as press releases provided evidence for disingenuous revenue projections, and indistinct business relationships. The case shows how both spam e-mails and false press releases are used as a fast, cheap and high impact means of disseminating false information.

The proposed service system uses data mining analysers to detect outliers and patterns of manipulation, utilising a number of financial indicators as inputs, similar to that in our earlier work [19]. More importantly, the proposed service system introduces the use of automated, linguistics-based text mining to extract key concepts from spam e-mails and press releases generated as part of ‘outing campaigns’. In this context, we can conclude that it would have been difficult to detect this manipulative scheme without these analysers. In fact, given the strategy followed by the insider traders in this case, it was not possible to raise early alarms of manipulation, and the fraud related actions were detected only when the actual manipulative trading occurred. Unstructured sources, such as spam-mails and forms, could be used to create proactive indicators and alerts of potential microcap manipulation cases, especially if these events are accompanied by changes in trading behaviour. Using on-demand analysis services from external providers could help investors and analysts to raise alarms earlier.

The key contribution of this work is to demonstrate the proposed approach and explain how it could be used within the existing fraud analysis processes, the extent to which the processes could be automated, the relationships with the other types of analysis and the role played by fraud analysts. Future work will expand the evaluation through the use of additional case studies, including the development of a comprehensive domain ontology which is expected to greatly improve the effectiveness of the proposed service system.

REFERENCES


Figure 14 Fraud Detection Engine Visualisation