

Data Mining Applications for Fraud Detection in Securities Market

Koosha Golmohammadi, Osmar R. Zaiane
Department of Computing Science
University of Alberta, Canada
Edmonton, Canada
{golmoham, zaiane}@ualberta.ca

This paper presents an overview of fraud detection in securities market as well as a comprehensive literature review of data mining methods that are used to address the issue. We identify the best practices that are based on data mining methods for detecting known fraudulent patterns and discovering new predatory strategies. Furthermore, we highlight the challenges faced in the development and implementation of data mining systems for detecting market manipulation in securities market and we provide recommendation for future research works accordingly.

Keywords: data mining, fraud detection, securities market, market manipulation, stocks

I. INTRODUCTION

Market capitalization exceeds \$1.8 trillion in Canada [1] and \$15 trillion in USA in 2010 (GDP of Canada and USA in 2010 are \$1.3 and \$14.6 trillion respectively). Protecting market participants from fraudulent practices and providing a fair and orderly market is a challenging task for regulators. Over 207 individuals from 100 companies were prosecuted in 2010 and this resulted in over \$120 million in fines, compensation and disgorgement in Canada. However, the effect of fraudulent activities in securities market and financial losses caused by such practices is far more than these numbers.

“Securities fraud broadly refers to deceptive practices in connection with the offer and sale of securities”. Securities fraud are divided into the following categories [2]:

- **High Yield Investment Fraud:** these schemes typically offer guaranteed returns on low-or-no-risk investments in securities instruments. Perpetrators take advantage of the investors’ trust and claim high returns to operate their funds. The most prevalent high yield investments appear in the form of: Pyramid Scheme, Ponzi schemes, Prime Bank Scheme, Advance Fee Fraud, Commodities Fraud (foreign currency exchange and precious metals fraud) and promissory notes.
- **Broker Embezzlement:** these schemes include broker unauthorized and illegal actions to gain profit from

the client’s investment. This may involve unauthorized trading or falsifying documents.

- **Late-Day Trading:** these schemes involve trading a security after market is closed.
- **Market Manipulation:** these schemes involve individuals, or a group of people attempting to interfere with a fair and orderly market to gain profit.

Market manipulation and price rigging remain the biggest concern of investors in today’s market, despite fast and strict responses from regulators and exchanges to market participants that pursue such practices [5]. Market manipulation is forbidden in Canada under Bill 30-46 [20] and in USA under Section 9(a)(2) of the Securities Exchange Act of 1934 [6]. In this paper we review data mining techniques for detecting and preventing market manipulation. We define market manipulation in securities as follows:

market manipulation involves intentional attempts to deceive investors by affecting or controlling the price of a security or interfering with the fair market to gain profit.

We review the English literature that was published after 2001 to identify (a) the best practices in developing data mining techniques (b) the challenges and issues in design and development, and (c) the proposals for future research, to detect market manipulation in securities market.

There are many challenges involved in developing data mining applications for fraud detection in securities market, including: massive datasets, accuracy, privacy, performance measures and complexity. The impacts on the market and the training of regulators are other issues that need to be addressed. In this paper we present the results of a comprehensive systematic literature review on data mining techniques for detecting fraudulent activities and market manipulation in securities market. We also highlight the challenges in developing data mining systems for market manipulation and identify directions for future research.

The remainder of this paper is organized as follows: In Section 2, we describe how the literature was searched and selected. In Section 3, we review numerous data mining techniques in the selected literature, which have been designed to detect market manipulation in securities market.

In Section 4, we highlight issues and challenges that designers and developers have in developing effective data mining methods for detection of market manipulation. We also make suggestions for future research directions. Section 5, concludes the paper.

II. METHODOLOGY

The coauthor of this paper collaborated with Cameron Schuler, who has extensive experience in securities market, forming an advisory panel. The advisory panel monitored progress of the literature review and provided guidance and input into the project (the advisory panel was not involved in reviewing papers).

A. Identifying Relevant Publications

We identified relevant papers in twofold: 1) papers that were extracted from computing science databases by searching relevant keywords and 2) papers that were recommended by our advisory panel. We searched English language literature on data mining applications in fraud detection in securities market that were published after 2001. In total, 261 papers were extracted from IEEE Xplorer, ACM Digital Library, and ISI Web of Knowledge databases using the following search terms exclusively or in combination, using the logical operators of “AND” and “OR”: “data mining”, “stock market”, and “fraud detection”. We also added 8 hand-selected articles that were provided by our advisory panel. After removing duplicates, a total of 205 articles remained to be screened for inclusion.

We screened the papers with the goal of identifying articles that address market manipulation in securities market. Papers were included for detailed review if they met the following two criteria:

- 1) The paper addresses fraud detection in securities market.
- 2) The proposed data mining methods or discussions in the paper are applicable to detecting market manipulation in securities market.

At the end of this process, 38 papers out of 205 papers met the criteria to be reviewed.

B. Annotating Publications

Each selected paper was carefully reviewed focusing on specific research questions in the study. In particular, we looked at the following issues about the context, method/approach and experiments, in each paper.

Context: we collected the information related to the specific target group envisioned by the paper (i.e. fraudulent activities that are associated with a party and/or fraudulent activities that are associated with a security), the type of market manipulation that is being addressed in the paper such as marking the close, wash trades and pre-arranged trading, cornering the market in a security, and insider trading.

Method/approach: we annotated papers based on the data mining approach (i.e. supervised, unsupervised and semi-

supervised learning), data mining method (eg. classification, outlier detection, social network analysis), data mining challenges, and issues.

Implementation: we collected information about the implementation level (eg. proof-of-concept, proposal supported by experimental results on real data), complexity, and scalability, for each reviewed paper.

Experiments: we annotated papers with different qualitative/quantitative evaluations that each paper provided to support its proposal. Specifically, we collected information on the evaluation method, data population, type of experiments/simulations, and qualitative/quantitative results on performance or comparison to other methods.

C. High-level Analysis

A program was developed to analyze 205 references that were extracted from computer science databases. This program extracts information such as the name of authors, publication year, keywords, URL, etc. This analysis gives a general picture of the most repeated, and arguably the most popular keywords and topics in the pool of extracted references. We used the same program to find the distribution of papers per year (see figure 1). This figure shows that there has been a significant growth in the number of publications related to market manipulation in securities market during the past few years. It should be mentioned, that the pool of papers that we have, was extracted in April 2012. We expect the number of publications in 2012, to continue the increasing trend.

We used voyeurtools¹, which is a web interface to TAPoRware (Text Analysis Portal for Research) - a suite of text-analysis services for researchers and scholars in the digital humanities² - to analyze the content of the 205 papers selected for review in this study.

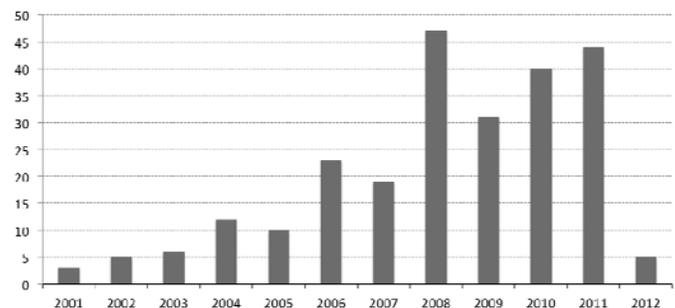


Figure 1. Distribution of 205 papers related to data mining for fraud detection in securities market per publication year

III. DISCUSSION

We organized our discussion on the specific contributions of the various publications we reviewed in

¹ <http://voyeurtools.org/>

² <http://taporware.ualberta.ca/>

detail in five categories, based on the data mining approach, goals, and input data:

- *Pattern Recognition*

Goal: detecting patterns that are similar to trends known to represent fraudulent activities. This goal can be pursued in two levels: a) detecting suspicious traders with fraudulent behavior, b) detecting securities that are associated with fraudulent activities (this is desirable, as regulators can freeze trading on such securities to ensure fair market for the participants. A real-time (on line) data mining process is essential in this case.

Data: a dataset that includes historical trading data for each trader account (in the case of “a”) or for each security (in the case of “b”) and a set of patterns/trends that are known to be fraud (labels).

- *Outlier Detection*

Goal: detecting observations that are inconsistent to the remainder of data. This can help in discovering unknown fraud patterns. Also, spikes can be detected effectively in this scenario according to the market conditions, instead of using a threshold to filter out spikes. Similar to the first group, this can be performed both in security and trader levels.

Data: historical trading information of traders or securities. Anomaly detection is generally achieved using clustering methods and doesn’t require labeled data.

- *Rule Induction*

Goal: extracting rules that can be inspected and used by auditors/regulators of securities market.

Data: historical trading information for each trader account as well as trader accounts that are labeled to be suspicious for fraudulent activity. It is also possible to extract rules that identify unknown patterns and irregularities using unlabeled data (unsupervised learning methods).

- *Social Network Analysis*

Goal: detecting trader accounts that collaborate to manipulate the market.

Data: historical trading information for each trader account. Here, additional data sources about employment history of traders and their relationships are required to be integrated into the dataset (eg. NASD uses a Central Registration Depository (CRD [4]) that maintains information of federally registered brokers).

- *Visualization*

Goal: producing visualizations that go beyond conventional charts enabling auditors to interact with

the market data and find malicious patterns. Visualization of the market data is both important for real-time monitoring and off-line investigations. Visualization can help auditors identify suspicious activities in securities and traders transactions.

Data: historical trading data or real-time stream of data about securities/traders transactions.

A. *Pattern Recognition*

Pattern recognition in securities market typically is performed using supervised learning methods on either daily or intraday data (tick data) where features include statistical averages and returns. Ogut et. al. used daily return, average of daily change and average of daily volatility of manipulated stocks and subtracted these numbers from the same parameters of the index [16]. This gives the deviation of manipulated stock from non-manipulated (index) and higher deviations indicates suspicious activities. The assumption in this work is price (consequently return), volume and volatility increases in the manipulation period and drops in the post-manipulation phase. The proposed method was tested using the dataset from Istanbul Stock Exchange (ISE) that was used in a related work to investigate the possibility of gaining profit at the expense of other investors by manipulating the market [35]. Experimental results show that ANN and SVM outperform multivariate statistics techniques (56% compared to 54%) with respect to sensitivity (which is more important in detecting price manipulation as they report correctly classified manipulated data points).

Diaz et. al. employed an "open-box" approach in application of data mining methods for detecting intraday price manipulation by mining financial variables, ratios and textual sources [33]. The case study was built based on stock market manipulation cases pursued by the US Securities and Exchange Commission (SEC) during 2009. Different sources of data that were combined to analyze over 100 million trades and 170 thousand quotes in this study include: profiling info (trading venues, market capitalization and betas), intraday trading info (price and volume within a year), and financial news and filing relations. First, using clustering algorithms, a training dataset is created (labeling hours of manipulation because SEC doesn't provide this information). Similar cases and Dow Jones Industrial Average (DJI) were used as un-manipulated samples. Second, tree generating classification methods (QUEST, C5.0 and C&RT) were used and tested using jack-knife and bootstrapping. Finally, the models were ranked using overall accuracy, measures of unequal importance, sensitivity and false positives per positives ratio. A set of rules were generated that could be inspected by securities investigators and be used to detect market manipulation. The results indicate:

- liquidity, returns and volatility are higher for the manipulated stocks than for the controlling sample
- although, it is possible to gain profit by manipulating the price of a security to deflate it’s price (short selling),

most market manipulators attempt to increase the stock price

- closing hours, quarter-ends and year-ends are "common preconditions for the manipulations"
- sudden jumps in volume of trading and the volatility of returns are followed by price manipulation in most cases

B. *Outlier Detection*

Anomaly and outlier refer to "observations in a data set that appear to be inconsistent with the remainder of that set of data". Ferdowsi et. al. applied Peer Group Analysis (PGA) to transactional data in stock market to detect outlier traders [23]. The dataset consists of three months of real data from Bangladesh stock market that is claimed to be an appropriate dataset as securities fraud mostly appears in emerging markets such as Bangladesh stock market. The data is represented using statistical variables (mean and variance) of buy and sell orders under fixed time periods. The npeer is set as a predefined parameter describing the number of objects in a peer group and controls the sensitivity of the model. A target object is decided a member of a peer group if members of the peer group are the most similar objects to the target object. After each time window (5 weeks) peer groups are summarized to identify the centroid of the peer group. Then, the distance of peer group members with the peer group's centroid is calculated using t-statistic and objects that deviate significantly from their peer are picked as outlier. Trader accounts that are associated with these objects are flagged as suspicious traders that suddenly behaved differently to their peer.

IBM Watson research center proposed an efficient method for detecting burst events in stock market [22]. First, burst is detected in financial data based on a variable threshold using the skewed property of data (exponential distribution), second, the bursts are indexed using Containment-Encoded Intervals (CEIs) for efficient storing and access in the database. This method can be used for fraud detection or identifying fraudulent behavior in the case of triggering fraud alarms in real time. The burst patterns of stock trading volume before and after 9/11 attack is investigated using the proposed approach and the experimental results confirm that the method is effective and efficient compared to B+tree.

C. *Rule Induction*

Data mining methods that generate rules are especially important because of the intrinsic features which rules provide for fraud detection in securities market. High transparency, easily comparable to existing regulatory/auditing rules and easily integrable to existing tools, are only a few features that make using rules very popular among auditors and investigators in securities market.

Abe. et. al. introduced an approach for rule induction by temporal mining of data. First, time series data is cleaned

(preprocessing) in two steps: a) the period of subsequence is determined, and b) the temporal pattern extraction is performed using a clustering algorithm (EM and K-means). Also relevant data attributes are selected manually or by using attribute selection algorithms. Second, a rule induction algorithm such as C4.5 [27], AQ15 [28] or Version Space [29] is used to produce if-then rules. An environment is developed using the proposed method and tested using a dataset that consists of temporal price data (price, volume, high, low, in total 13 trend index) of 9 stocks from Japan's stock market from January 5, 2006 to May 31, 2006. The buy/sell decisions on each stock is determined using the clustering method and is used for testing on a different stock. Experimental results show that the introduced method for pattern extraction is promising as it outperforms the baseline.

A crucial issue in rule induction methods is identifying effective rules from the set of generated rules. There are numerous objective rule interestingness measures that can be used for this purpose. An extensive experiment comparing over 70 different objective measures to describe rule interestingness using a dataset in healthcare identified Recall [24], Jaccard [25], Kappa [25], Collective Strength (CST) [25], X2-M [24] and Peculiarity [26] as the most effective objective measures. However, such ranking may be different in experiments on financial data and to the best of our knowledge there hasn't been a work that compares objective measures for rule interestingness on financial data.

D. *Social Network Analysis*

Traditional data mining methods (e.g. classification, clustering, association rules) often consider samples as independent data points [30]. However, these methods cannot leverage the relationship between samples in datasets that are richly structured and mostly heterogeneous. Such structured data can be represented in the form of a Social Network where nodes correspond to data samples (i.e. objects/individuals), and edges represent relationships and dependencies between objects. Mapping, understanding, analyzing and measuring interactions across such a network is known as Social Network Analysis (SNA). Using SNA to find correlations that indicate fraud in securities market begins with transforming the market events to a graph (preprocessing). The most interesting application of SNA in securities market fraud is detecting brokers that collaborate to: a) inflate/deflate the price of a security by putting prearranged orders with other brokers and manipulating the volume, b) move stocks between account for tax reasons, and c) get credibility in the market with high number of transactions. Blume et. al. combined SNA and interactive visualization to identify malicious accounts in an exchange. Authors designed indicators of fraudulent activities (based on the textual description of typical fraud cases) that can be detected using SNA:

- Circular trading: consistently buying and selling more or less the same volume of a stock
- Primary-Secondary Indicator: marks accounts buying low and selling high. Network centrality can

help to find the primary account; a function of f is calculated for every vertex representing the size of the account and comparing the price of the transaction with average price (in the past c transactions)

- Prominent edge indicator: transferring stocks from one account to another which happens when an edge (transaction) between two vertices appears several times

SNA provides many algorithms that are effective in finding collaborative efforts to manipulate market as well as methods for monitoring interactions of traders in the market.

E. Visualization

Securities market investigators use different charts and figures to monitor the market. However, in our discussions with Canadian securities market auditors and regulators we found great interests in finding data visualization techniques that are beyond charts/tables which permit one to see the patterns within the data or other information not readily discernable. Stockare is a visual analytics framework for stock market [32], which combines a 3D Treemap for market surveillance, and a behavior-driven visualization using SNA for monitoring the brokers' activities. In the 3D visualization each cell represents a security, the size of a cell is proportional to the market capitalization and the color code of a cell indicates the change in the price (eg. green for increase and red for decrease in the price). The 3D visualization provides a tool for the real-time monitoring (15 minutes delay) of raw trading flow (price and volume). Trading details are compared to a set of parameters and an alert is raised if they are out of range. Analysis of the trading network aims to "reveal the social structure among traders" and "identify suspected trading patterns". Nodes represent traders, the area around each node represents the trading value, and directional edges indicate the flow and weight of trades/exchanges. A database of malicious trading patterns is used as a reference to compare with events in the trading network and identify suspicious activities.

Liquidity, returns and volatility are higher for the manipulated stocks, therefore, charting these parameters in parallel with the same time alignment helps regulators in identifying suspicious patterns and trends [33]. Isolated jumps in liquidity can indicate suspicious trades when returns are within the normal ranges.

IV. CHALLENGES AND FUTURE PERSPECTIVE

Data mining methods support fraud detection in securities market and experimental results in literature are encouraging. However, there are many challenges in designing and developing data mining methods for detecting price manipulation in securities market including heterogeneous datasets, privacy, performance and legal consequences. These issues should direct future research and designs.

A. High Frequency Trading

High Frequency Trading (HFT) is an evolution of securities market that adopts sophisticated algorithms to automatically analyze and react to market data in milliseconds. It is estimated that HFT accounts for 35% of the stock market trades in Canada and 70% of the stock trades in USA [18]. HFT strategies can be divided into five categories:

- 1) **Passive Rebate Arbitrage:** providing liquidity and receiving incentives from exchanges similar to market makers that is one of the major HFT strategies
- 2) **Latency Arbitrage:** making profit through buying and selling the same security between domiciles (i.e. Inter-listed Arbitrage) or between domestic marketplaces (i.e. Intra-listed Arbitrage)
- 3) **Information Arbitrage:** making profit through buying and selling the same security at a higher price by taking advantage of "mispricing between the various forms of a tradable index"
- 4) **Statistical Arbitrage:** applying statistical methods to detect trading patterns for *relative value trading*
- 5) **Market Structure Trading:** making profit through trading opportunities that are a consequence of the new market structure such as flash orders, locked markets, and dark pools

While HFT supports liquidity and contributes to price formation in the market, it might have negative impacts in adverse market conditions. Regulators have been considering trading obligations and supervision on HFT especially after the May 6 flash crash [34]. Both growth and impact of HFT in stock market have brought great interests of industry and thus requires regulators to establish an environment to support fair and orderly trading market. Unlike traditional trading, HFT is not subject to significant trading obligations and there is very little public information regarding fraudulent patterns and activities of HFT systems. Data mining techniques can be employed to identify fraudulent activities and predatory strategies in HFT.

B. Massive Data

The datasets in securities market are huge. There are over 2700 securities listed in NASDAQ and SuperMontage (NASDAQ's trading platform) facilitates more than 5000 transactions per second. Similarly, the number of transactions show a significant growth from 250,000 to 5 million orders per day a few years ago, to 700,000 to 40 million orders each day³. The rate of growth in the amount of data is rapidly increasing due to changes in trading strategies by both buy-side and sell-side firms. Thus, there is extra pressure on industry to accommodate faster trading

³ Canadian Securities Administrators (CSA), market regulation service report 2008

systems and on regulatory organizations to adapt to new strategies.

C. Different forms of data

The data in securities market comes from different resources and in different forms such as news data, analytical data (level I and level II data) and fundamental data (financial reports and filings). We categorized the data in securities market to two groups:

- **Unstructured data**
 - News and Financial Events (eg. Factiva⁴)
 - Stock-chat message board (eg. stocktraderchat⁵, yahoo message board⁶)
- **Structured data**
 - Trading data (eg. Trade And Quote (TAQ) from NASDAQ⁷)
 - Stock Analytics
 - Companies financial information (COMPUSTAT⁸)
 - Companies Insider Activities (eg. Thomson Reuters Insider Filings Data Feed (IFDF))

Heterogeneous datasets and integrating data from different sources makes both data preparation and learning phase of the data mining process challenging.

D. Unlabeled data

Labeled data for fraud detection in securities market is very rare because (a) labeling data is very costly and typically requires investigation of auditors, and (b) the number of positive samples (fraud cases) constitute a tiny percentage of the total number of samples. Koscsis et. al. used MDP to generate synthetic samples (assuming there are very few positive samples) and used frequency that a player abates from the optimal policy as features to train the classifier in the modeling[36].

E. Performance Measure

Misclassification costs are unequal in fraud detection because false negatives are more costly. In other words, missing a fraud case by predicting it to be not-fraud, hurts performance of the method more than including a not-fraud case by predicting it to be fraud. The most effective performance measures for fraud detection are listed below based on the learning approach:

- **Supervised Learning:** threshold, ordering, and probability metrics are effective performance measures for evaluating supervised learning methods for fraud detection [7]. According to our studies the

most effective metrics include: AMOC [9]: “Activity Monitoring Operating Characteristic (AMOC) (average score versus false alarm rate), Receiver Operating Characteristic (ROC) analysis (true positive rate versus false positive rate), maximizing Area under the Receiver Operating Curve (AUC), minimizing cross entropy (CXE) [10], minimizing Brier score [10] and mean squared error of predictions.

- **Semi-supervised Learning:** the most effective performance measures for fraud detection using semi-supervised learning methods (eg. anomaly detection) include: entropy, conditional entropy, relative conditional entropy, information gain and information cost [11].
- **Unsupervised Learning:** application of Hellinger and logarithmic scores [12] and t-statistic [14] resulted in higher performances when using data mining methods that are based on unsupervised learning approaches.

V. CONCLUSION

The significant growth of the capital market due to increasing interests in investing in securities market requires regulatory organizations to expand their efforts to ensure a fair and orderly market for the participants. Data mining methods are effective in detecting different types of fraud in securities market. In this paper we reviewed the most significant data mining methods that are applied to monitoring securities market and detecting market manipulation. We highlighted numerous challenges that are involved in developing data mining methods for detection and prevention of fraudulent activities in securities market. Some of the challenges in designing and developing data mining methods include massive datasets, different sources and forms of data and using appropriate performance measures to evaluate the method. We also provided directions for new research in this field.

ACKNOWLEDGMENTS

The authors wish to thank Cameron Schuler for his expert advice in securities market. We also acknowledge the financial support from Natural Science and Engineering Research Council (NSERC), Alberta Innovates Centre for Machine Learning (AICML) and Alberta Innovates Graduate Student Scholarship (AITF).

REFERENCES

- [1] Canadian Securities Administrators (CSA) 2010 report, last access March 7, 2012 at: http://www.osc.gov.on.ca/documents/en/About/csa_20110222_csa-rpt-enf-2010.pdf
- [2] FBI report 2010-2011, last access March 7, 2012 at: <http://www.fbi.gov/stats-services/publications/financial-crimes-report-2010-2011>

4 global.factiva.com/

5 <http://stocktraderschat.com/>

6 <http://biz.yahoo.com/promo/mbbeta.html>

7 <https://www.nasdaqdod.com/Products/Catalog.aspx>

8 <http://www.compustat.com/>

- [3] Alberta Securities Commission (ASC) 2010 report, last access March 7, 2012 at:
<http://www.albertasecurities.com/news/ASC%20Publications/6116/2010%20ASC%20Annual%20Report.pdf>
- [4] J. Neville, D. Jensen, J. Komoroske, K. Palmer, H. Goldberg. "Using Relational Knowledge Discovery to Prevent Securities Fraud", 11th ACM SIGKDD International Conference on Knowledge Discovery in Data mining series, pp. 449-458, 2005
- [5] <http://economictimes.indiatimes.com/markets/stocks/market-news/market-manipulation-continues-to-be-the-biggest-concerns-for-investors/articleshow/12076298.cms>, 29 Feb. 2012, last access March 7, 2012
- [6] www.sec.gov/about/laws/sea34.pdf, last accessed March 12, 2012
- [7] Phua, V. Lee, K. Smith, and R. Gayler. A comprehensive survey of data mining-based fraud detection research. *Artificial Intelligence Review*, 2005
- [8] Caruana, R. & Niculescu-Mizil, A. (2004). Data Mining in Metric Space: An Empirical Analysis of Supervised Learning Performance Criteria. *Proc. of SIGKDD04*, 69-78
- [9] T. Fawcett, F. Provost. "Activity monitoring: Noticing Interesting Changes in Behavior", In proceedings of SIGKDD99, pp. 53-62, 1999
- [10] S. Viaene, R. Derrig, G. Dedene. "A Case Study of Applying Boosting Naive Bayes to Claim Fraud Diagnosis", *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, pp. 612- 620, 2004
- [11] W. Lee, D. Xiang. "Information-theoretic Measures for Anomaly Detection", In proceedings of 2001 IEEE Symposium on Security and Privacy, pp. 130-143, 2001
- [12] K. Yamanishi, J. Takeuchi, G. Williams, P. Milne. "On-Line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms", *Journal of Data Mining and Knowledge Discovery*, vol. 8, pp. 275-300, 2004
- [13] P. Burge, J. Shawe-Taylor. "An Unsupervised Neural Network Approach to Profiling the Behaviour of Mobile Phone Users for Use in Fraud Detection", *Journal of Parallel and Distributed Computing*, vol. 61, pp. 915-925, 2001
- [14] R. Bolton, D. Hand. "Unsupervised Profiling Methods for Fraud Detection", In roceeding of Credit Scoring and Credit Control VII, 2001
- [15] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth. "CRISP-DM 1.0 Step-by-step data mining guide", The CRISP-DM consortium, 2000
- [16] H. Ogut, M. M. Doganay, R. Aktas. "Detecting stock-price manipulation in an emerging market: The case of Turkey", *Journal of Expert Systems with Applications*, vol. 36, pp. 11944-11949, 2009
- [17] IIROC Market Regulation Rervices, Annual Report 2008, available at
http://sdocs.iroc.ca/English/Documents/2008/AE98E83-B-C25F-48F4-B4BD-06D33811772E_en.pdf, last accessed on March 15, 2012
- [18] K. Swinburne. "On the Regulation of Trading in Financial Instruments: Dark pools & HFT", report for the European Union, 2010
- [19] L. Koscis, A. Gyoergy. "Fraud Detection by Generating Positive Samples for Classification from Unlabeled Data", *ICDM 2010*, Israel, 2010
- [20] http://www.parl.gc.ca/About/Parliament/LegislativeSummaries/bills_ls.asp?Language=E&ls=C46&Mode=1&Parl=37&Ses=2&source=library_prb, last accessed on April 24, 2012
- [21] V. Barnett and T. Lewis. "Outliers in Statistical Data", John Wiley & Sons, 1994
- [22] M. Vlachos, K. Wu, S. Chen, P.S. Yu. "Correlating burst events on streaming stock market data", *Journal of Data Mining and Knowledge Discovery*, vol. 16, pp. 109-133, 2008
- [23] Z. Ferdousi, A. Maeda. "Unsupervised outlier detection in time series data", 22nd International Conference on Data Engineering Workshops, pp. 51-56, 2006
- [24] M. Ohsaki, Y. Sato, H. Yokoi, T. Yamaguchi. "Investigation of Rule Interestingness in Medical Data Mining", *Lecture Notes in Computer Science*, vol. 3430, pp. 174-189, 2005
- [25] P. N. Tan, V. Kumar, J. Srivastava. "Selecting the Right Interestingness Measure for Association Patterns", *KDD*, pp. 32-41, 2002
- [26] N. Zhong, Y. Y. Yao, M. Ohshima. "Peculiarity Oriented Multi-Database Mining", *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, pp. 952-960, 2003
- [27] J. R. Quinlan. "Programs for Machine Learning", Morgan Kaufmann, 1992
- [28] R. Michalski, I. Mozetic, J. Hong, and N. Lavrac. "The AQ15 Inductive Learning System: An Overview and Experiments", *Reports of Machine Learning and Inference Laboratory, MLI-86-6*, George Mason University, 1986
- [29] T. M. Mitchell. "Generalization as Search", *Artificial Intelligence*, vol. 18, pp. 203-226. 1982
- [30] L. Getoor, C. P. Diehl. "Link mining: a survey", *ACM SIGKDD Explorations*, vol. 7, pp. 3-12, 2005
- [31] M. Blume, W. Christof. "Using Network Analysis for Fraud Detection in Electronic Markets", *Journal of Information Management and Market Engineering*, vol. 4, pp. 101-112, 2006
- [32] M.L. Huang, J. Liang, V. Quang, N. Vinh. A Visualization Approach for Frauds Detection in Financial Market", 13th International Conference of Information Visualization, pp. 197-202, 2009
- [33] D. Diaz, B. Theodoulidis, P. Sampaio. "Analysis of stock market manipulations using knowledge discovery techniques applied to intraday trade prices", *Journal of Expert Systems with Applications*, vol. 38, pp. 12757-12771, 2011

- [34] Findings Regarding the Market Events of May 6, 2010, Report of the staffs of the CFTC and SEC to the joint advisory committee on emerging regulatory issues.
- [35] R. Aktas, M. Doganay. "Stock-price manipulation in the Istanbul stock exchange", *Journal of Eurasian Review of Economics and Finance*, vol. 2, pp. 21–28, 2006
- [36] L. Kocsis, A. Gyorgy. "Fraud Detection by Generating Positive Samples for Classification from Unlabeled Data", *ICML 2010 Workshop on Machine Learning and Games*, Israel, 2010