UNIVERSIDAD DE OVIEDO

MASTER IN SOFT COMPUTING
AND INTELLIGENT DATA ANALYSIS

PROYECTO FIN DE MASTER
MASTER PROJECT

LUIS ALVARO BARRIENTOS NIÑO DE GUZMÁN
JULY 2012

UNIVERSIDAD DE OVIEDO

MASTER IN SOFT COMPUTING
AND INTELLIGENT DATA ANALYSIS

PROYECTO FIN DE MASTER
MASTER PROJECT

INSIDER TRADING SEQUENTIAL PATTERN MINING (INTRASPAM)

LUIS ALVARO BARRIENTOS NIÑO DE GUZMÁN
JULY 2012

TUTOR/ADVISOR:
CHRISTIAN BORGELT

Declaration of Authorship

I, Luis Alvaro Barrientos, declare that this work titled, 'Insider Trading Sequential Pattern Mining (INTRASPAM)' and the content presented in it is my own. I confirm that:

- This work was done wholly while in candidature for a master's degree at this University.
- Where I have consulted the published work of others, this is always clearly attributed.
- I have acknowledged all main sources of help.

Signed: Alvaro Barrientos

Date: 18-07-2012

"The quickest way to double your money is to fold it over and put it back in your pocket."

Will Rogers

UNIVERSIDAD DE OVIEDO

Abstract

European Centre for Soft Computing

Master in Soft Computing and Intelligent Data Analysis

INSIDER TRADING SEQUENTIAL PATTERN MINING (INTRASPAM)

Insider trading filings are a source of publicly available information available to anyone by the United States Securities and Exchange Commission (SEC). These filings are mandatory for all corporate executives as well as for major shareholders willing to trade shares of their own company. The fact that financial markets allow people to profit from knowledge about events that affect a company, provokes an interest in the behavior of those agents and their trading movements given that they have an information unavailable to the public. This work is aimed at discovering sequential associations within the insider trading information stored at the United States SEC's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system. In order to seek relationships from a potentially vast amount of information having a temporal ordering, the information is treated using a sequential pattern mining algorithm able to handle time constraints such as cSPADE. [1] The needed infrastructure is created under a R environment [2] along with the arulesSequences [3] package. Finally, frequent sequential patterns are obtained before and after the Lehman Brothers bankruptcy filing on September 15, 2008, in order to look for dissimilarities due to this event.

Acknowledgements

I would like to thank my advisor Prof. Christian Borgelt who has not only introduced me to Data Mining, but also encouraged me to follow my ideas while sharing valuable insights I would've not grasped otherwise. I would also like to thank and acknowledge the European Centre for Soft Computing Scholarship Programme for materializing the opportunity to study this Master. I would also like to thank my family for their continuous support no matter where I am or what I do, nothing could have been possible without their sacrifices.

Contents

List of Figures

List of Tables

Abbreviations

CIK	**C**entral **I**ndex **K**ey
cSPADE	**c**onstrained **S**equential **PA**ttern **D**iscovery using **E**quivalent Class
EDGAR	**E**lectronic **D**ata **G**athering, **A**nalysis, and **R**etrieval
EID	**E**vent **ID**entifier
FTP	**F**ile **T**ransfer **P**rotocol
NA	**N**ot **A**vailable
SEC	**S**ecurities and **E**xchange **C**ommission
SIC	**S**tandard **I**ndustrial **C**lassification
SID	**S**equence **ID**entifier
URL	**U**niform **R**esource **L**ocator
XML	e**X**tensible **M**arkup **L**anguage

Chapter 1

Introduction

1.1 Insider Trading

Nowadays and even before the start of the latest major financial crisis, one of the most controversial and analyzed entities is the stock market. In order to trade a company stock there must be an agreement on how much money the underlying fraction of ownership is worth given the past, present, and future of that company. This valuation is dynamically changing as market participants -which include individual retail investors, institutional investors, and also publicly traded corporations trading in their own shares- are constantly evaluating if that price agreement might be higher or lower in the future given the current information available.

It seems logical to assume that the more detailed, accurate, and relevant information a market participant has about a company, the greater the possibilities of profiting from that company's stock that participant has. [4] If we follow this reasoning, it can be inferred that the people having at their disposal greater amounts of material information are the ones involved in the most sensitive aspects of a company. Therefore, in countries like the United States, corporate insiders are defined as the directors, company's officers, and any major owner of more than ten percent of a company's equity securities. As insiders, their trading activity in their company's own stock is regulated by the SEC under Section 10(b) of the Securities Exchange Act of 1934. This regulation prohibits those trades based on material non-public information and would be considered as misappropriation of corporate information. [5] For example, given the knowledge of an insider prior to earnings announcement, it would be considered a violation of the law if trading activity took place beforehand, this is known as active trading. On the other hand, if trading takes place after earnings announcement it is allowed. Moreover, law does not penalize insiders whose benefit come from non-public information when they use it to avoid buying or selling stocks, because it would be considered passive trading.

Thus trading abstention is not expected at all from insider participants, actually insider trading takes place besides the constrained passive way, but also as active trading despite the risk of being caught as detection or enforcement by the SEC is seen as unlikely. [6] It is also known from past cases that insider trading is highly correlated with the type of event released. Certain news are more likely to be associated to active trading events, over 85% of past insider trading cases were based on material news of either product announcements, earnings announcements, regulatory approvals or denials, mergers and acquisitions, or research reports. [7] Supported by several studies, it has been concluded that corporate insiders who trade shares of their own companies earn abnormal returns. [8–12] This possibility has created a market for information on insider trading. If insiders trade on private information, then replicate insiders' trading patterns and behavior is highly desirable, either by examining insider trading statistics and/or independently discovering of common occurrences within their behavior.

1.2 Data Mining

Discovering associations is one of the tasks which characterize data mining. As the verb itself, mining or seeking dependencies from potentially huge amounts of data is a labor that calls for an efficient and statistically relevant approach. Its aim is to seek automatically for dependencies from vast amounts of data. The possible outcome is interpreted as: *"If A occurs in the data then B also occurs"* This kind of relations are known as association rules [13], which will be significant only if they occur in the data frequently enough. Even though several algorithms have been successfully approached in [13–19] to obtain association rules, there are several real-world domains where data is generated along with a temporal component. This kind of information is also exploitable and together compose sequences which might discloses patterns in a sequence database.

Sequential pattern mining [20–22], proved to be an active research topic having a broad application area. It is currently being applied to the analysis of customer purchase patterns, web usage patterns, DNA sequences, time-related processes, among many others. The underlying sequential associations or sequential patterns in these kinds of domains are similar to the ones obtained in traditional association rules, but in the following way: "When A occurs, B occurs within certain time" For some applications it may also be relevant to take into account the gaps among and within sequential patterns as well. Besides this sensitiveness, it might happen that exact inflexible sequences are not desirable for specific applications where there is an intra or inter sequence proportion of noise expected. Therefore, adding time constraints for better mining results is an important problem. In order to enhance the semantics of sequence discovery, time constraints such as minimum gap, maximum gap and a sliding time-window were specified in [23] for the first time. A minimum gap allows exclusion of undesired patterns, meanwhile a maximum gap helps to filter out if the time gap between two adjacent transactions in a pattern is too long for the purpose. The sliding window may allow transactions to be handled as one, providing a continuous context.

1.3 Motivation

By the time this work took place, approximately half of a decade has passed since the last major financial crisis started in the United States. It is easily remembered as an inflection point for different reasons, for a lot of people.

The research community has been trying to come up with models of the stock market behavior, and probably every single related variable related to this application area. Yet another event within this financial crisis such as the 2010 flash-crash[1] reflects the increasing presence of algorithms and HFT[2] technology have in the stock market behavior nowadays. So, in this thesis a frequent pattern discovery process is conducted in an area which is usually overlooked by machine learning researchers as is the case with insider trading, and to our best knowledge no work in the literature has been done following this kind of approach. A literature survey [24] shows that most of the approaches involving both data mining and insider trading is almost completely focused on financial crime detection or fraud. The main difference lies in finding common but frequent sequential patterns as the objective of the study, meanwhile fraud detection attempts to find sequences that are considered irregular. Moreover, another objective of this thesis is to compare all the temporal associations among the frequent patterns which are greater than a fixed minimum support, before and after one of the turning points in the financial crisis such as the bankruptcy of Lehman Brothers on September 15, 2008, the largest one in United States history so far.

[1]The Flash Crash was a United States stock market crash on Thursday May 6, 2010 in which the Dow Jones Industrial Average plunged about 1000 pointsor about nine percentonly to recover those losses within minutes.

[2]High-frequency trading is the use of sophisticated technological tools to trade securities like stocks or options.

1.4 Organization

The Thesis is organized in five chapters. Chapter 1 provides a brief introduction to what insider trading is, the data mining area, and describes the motivation of this thesis. Chapter 2 reviews relevant background theory in the area of sequential pattern data mining, and the cSPADE algorithm. Chapter 3 describes the data source, gathering, exploration, and pre-processing required, as well as an analysis of the overall data. Chapter 4 comprises a series of experiments along with their correspondent obtained sequential patterns. Finally, Chapter 5 is about the discussion of Conclusions and future work.

Chapter 2

Background Theory

2.1 Basic Notions

Data mining is the process of finding patterns, associations, and statistically significant structures from highly populated databases. Besides the fact these structures must become reliable predictions, they are required to achieve understandable descriptions as well. How a prediction decision is arrived at, might not be relevant. Meanwhile, its description is appreciated by the clarity and simplicity of the model describing the problem. To be considered as significant, ideally, a rule should also remain valid in future new data samples, be new to experts, lead to useful actions, and provide a new insight on the problem. A general formulation of the basic notions of frequent pattern mining is introduced as it is presented in [25].

Before data mining algorithms can be used, a target data set must be assembled. Target data set is usually represented as a table, where each row of the table is a record, and each column is an attribute of the data. An attribute is an item that describes an aspect of the data sample. Given such a table, the value stored in jth column of the ith row is the value of the jth attribute of the ith data sample. A *feature* f and its value v in a sample are also referred to as an *item*.

A set of items is then called an *itemset* and can be written like $\{f_1 = v_1, ..., f_n = v_n\}$ for an itemset containing features $f_1, ..., f_n$ and associated values $v_1, ..., v_n$. Given such an itemset I, we denote by $[i]_{f_i}$ the value of its feature f_i. An itemset can also be represented as a feature vector $\langle v_1, ..., v_n \rangle$, where the value of the feature f_i is kept in the ith position of the vector. An itemset containing k items is called a k-itemset, where the number k is the cardinality of the itemset.

An itemset is also called a *transaction* when we let $B = \{i_1, i_2, ..., i_m\}$ be a set of m binary features called *items*. Note that transactions contain only those items whose feature values are 1 and not those whose values are 0. This set is also called the *item base*. Any subset $I \subseteq B$ is called an *item set*.

Let $T = (t_1, t_2, ..., t_n)$ with $\forall k,\ 1 \leq k \leq n : t_k \subseteq B$ be a vector of transactions over B. This vector is called the *transaction database*. Every transaction is an item set, but some items may not appear in T. Transactions are not necessarily unique, and it may be $t_j = t_k$ for $j = k$, and the item base B may not be explicitly given, but implicitly given as $B = \bigcup_{k=1}^{n} t_k$.

We say a transaction $t \in T$ covers an *item set* I, or *item set* I is *contained in* a transaction $t \in T$ iff $I \in t$. The concept of *cover* of an item set can be thought of as ”a vector of transactions that covers it” [25] Formally, the set $K_T(I) = \{k \in \{1, ..., n\} | I \subseteq t_k\}$ is called the *cover* of I w.r.t. T.

Definition 2.1. Let a dataset T of transactions and an itemset I be given. We denote the dataset cardinality by $|T|$. Then the **relative support** of I in T, denoted $\sigma_T(I)$, is the ratio of transactions in T that contain I, that is:

$$\sigma_T(I) = \frac{|K_T(I)|}{|T|}$$

Equivalently, the **absolute support** of I in T, is denoted as:

$$s_T(I) = |K_T(I)|$$

Given a **minimum support** $\sigma_{min} \in \Re, 0 \leq \sigma_{min} \leq 1$ or $s_{min} \in \aleph, 0 \leq s_{min} \leq |T|$, we can obtain the set of *frequent item sets* $\Phi_T(\sigma_{min}) = \{I \subseteq B | \sigma_T(I) \geq \sigma_{min}\}$ or (equivalently) $F_T(s_{min}) = \{I \subseteq B | s_T(I) \geq s_{min}\}$.

If we have an item set $J \subseteq B$, and the following support property is considered:

$$\forall I : \forall J \supseteq I : K_T(J) \subseteq K_T(I)$$

Which holds since:

$$\forall t : \forall I : \forall J \supseteq I : J \subseteq t \rightarrow I \subseteq t$$

It means each additional item is another condition a transaction has to satisfy, and the ones that do not satisfy it are removed from the cover. Then if an item set is extended, its support cannot increase or is *anti-monotone*. It is followed that:

$$\forall s_{min} : \forall I : \forall J \supseteq I : s_T(I) < s_{min} \rightarrow s_T(J) < s_{min}$$

This property is known as the **Apriori Property**. [13] Given a threshold s_{min}, an itemset I is said to be frequent in a dataset T if $s_T(I) \geq s_{min}$. Moreover, a frequent itemset I is said to be maximal if none of its proper supersets is frequent. This also implies that all subsets of a maximal frequent itemset are frequent, and closed if none of its supersets has the same frequency.

By exploiting these properties, it is now possible to obtain association rules from a dataset. An association rule $X \Rightarrow Y$ can be interpreted as:

"If X is observed in data sample T, then Y is also likely to be observed in T".

There are two important properties associated with rules. The first property is the support of the rule. The second property is the confidence of the rule.

Definition 2.2. The support of the rule $X \Rightarrow Y$ in a dataset T is defined as the percentage of transactions in T that contain $X \cup Y$.

$$s_T(X \Rightarrow Y) = s_T(X \cup Y)$$

Definition 2.3. The confidence of the rule $X \Rightarrow Y$ in a dataset T is defined as the percentage of transactions in T containing X that also contain Y, alsocan be seen as an estimate of $P(Y|X)$.

$$c_T(X \Rightarrow Y) = \frac{s_T(X \cup \ Y)}{s_T(X)} = \frac{\sigma_T(X \cup Y)}{\sigma_T(X)}$$

Support and confidence are two properties for determining if a rule is interesting. However, these may not show whether a rule is really interesting for a particular application. Setting minimum support, and confidence is usully application-dependant.

2.2 Sequential Pattern Mining

Sequential mining is a problem closely related to association rules mining. The difference lies in the data sample which is usually a list of events ordered in time. For association rules, each row in the input table represents a single data sample, where each row is the entire transaction. For sequence mining, data sample—called a sequence—is split across multiple consecutive rows in the input table, where each row represents just one event of the sequence. [26]

A sequence s, denoted by $< \alpha_1 \to \cdots \to \alpha_n >$, is an ordered list of n elements where each element α_i is an itemset. Without loss of generality, it is assumed that the items in an element are in lexicographic order.

The length of a sequence s is the total number of items in all the elements in s, denoted by $|s|$. Sequence s is a k-sequence if $|s| = k$. Given a dataset of sequences T containing $|T|$ data sequences, data sequence s has unique sequence identifier sid and is represented by $< \alpha_1^{t_1} \to \cdots \to \alpha_n^{t_m} >$, where element α_i happens at time $t_i, t_1 < \cdots < t_m$.

Definition 2.4. A sequence of events $< \alpha_1 \to \cdots \to \alpha_n >$ is contained in a sequence $< \beta_1 \to \cdots \to \beta_m >$ if there is a mapping $\varphi : \{1, ..., n\} \mapsto \{1, ..., m\}$ such that (1) for $1 \leq i \leq n$, $\alpha_i \subseteq \beta_{\varphi(i)}$; and (2) for $1 \leq i \leq j \leq n$, $\varphi(i) \leq \varphi(j)$. A sequence is maximal if it is not contained in other sequences. Then it is written $s_1 \subseteq s_2$ if the sequence s_1 is contained in the sequence s_2.

The notion of support is extended to sequences as follow:

Definition 2.5. The support of a sequence s in dataset of sequences T is the percentage of sequences in T that contain s. That is,

$$\sigma_T(s) = \frac{|\{s' \in T \mid s \subseteq s'\}|}{|T|}$$

A sequence of events $< \alpha_1 \to \cdots \to \alpha_n >$ can generate $n-1$ rules of the form $X \Rightarrow Y$, $< \alpha_1 > \Rightarrow < \alpha_2 \to \cdots \to \alpha_n >$, $< \alpha_1 \to \alpha_2 > \Rightarrow < \alpha_3 \to \cdots \to \alpha_n >$, ..., and $< \alpha_1 \to \cdots \to \alpha_{n-1} > \Rightarrow < \alpha_n >$. The notions of support and confidence on rules can then be defined in a manner analogous to itemset association rules.

Definition 2.6. The support of the rule $X \Rightarrow Y$ in a dataset T of sequences is defined as the percentage of sequences in T that contain $X \to Y$.

$$\sigma_T(X \Rightarrow Y) = \frac{|\{s' \in T \mid X \to Y \subseteq s'\}|}{|T|}$$

Definition 2.7. The confidence of the rule $X \Rightarrow Y$ in a dataset T of sequences is defined as the percentage of sequences in T containing X that also contain $X \to Y$.

$$c_T(X \Rightarrow Y) = \frac{\sigma_T(X \to Y)}{\sigma_T(X)} = \frac{|\{s' \in T \mid X \to Y \subseteq s'\}|}{|\{s' \in T \mid X \subseteq s'\}|}$$

A sequence s in the sequence dataset T is a time-constrained sequential pattern if its support in T satisfies user-specified minimum support σ_{min}. The support calculation also has to satisfy three time-constraints $maxgap$, $mingap$,and win.

- *maxgap*: It is the maximum allowed time difference between the latest occurrence of an event in an event-set and the earliest occurrence of an event in its immediately preceding event-set
- *mingap*: It is the minimum required time difference between the earliest occurrence of an event in an event-set and the latest occurrence of an event in its immediately preceding event-set
- *win*: It is the maximum allowed time difference between the latest and earliest occurrence of events in the entire sequence

Definition 2.8. Data sequence $s =< \alpha_1^{t_1} \to \cdots \to \alpha_n^{t_m} >$ contains a sequence $s' =< \beta_1 \to \cdots \to \beta_w >$ if there exist integers $p_1, q_1, p_2, q_2, \cdots, p_w, q_w$ and $1 \leq p_1 \leq q_1 < p_2 \leq q_2 < \cdots < p_w \leq q_w \leq n$ such that the following conditions hold:

$$\beta_i \subseteq (\alpha_{pi} \cup \cdots \cup \alpha_{qi}), 1 \leq i \leq w$$

$$t_{qi} - t_{pi} \leq win, 1 \leq i \leq w$$

$$t_{qi} - t_{pi} - 1 \leq maxgap, 2 \leq i \leq w$$

$$t_{pi} - t_{qi} - 1 \leq mingap, 2 \leq i \leq w$$

It is assumed that t_j, $mingap$, $maxgap$, win are all positive integers, $maxgap \geq mingap \geq 1$. Data sequence s is a supersequence of sequence s' (s' is a subsequence of s) if s contains s'. When $mingap$ is the same as $maxgap$, the time constraint is additionally called exact gap. Common sequential pattern mining without time constraints is a special case by setting $mingap = 1$, $maxgap = \infty$, $win = 0$.

2.3 cSPADE Algorithm

The cSPADE (constrained Sequential PAttern Discovery using Equivalence classes) algorithm [1] is designed for transaction data, using efficient lattice search techniques and simple joins. All sequences are discovered with only three passes over the database by using equivalence prefix classes, which allow the search space to be decomposed into smaller subproblems and fit in main memory. It is a straightforward extention of the earlier SPADE algorithm [27], the difference lies in the integration of constraints like length, width, and duration limitations on the sequences.

In this approach, the sequence database is represented as a vertical occurrence list database (id-list) denoted by $\zeta(X)$ for an event X, in which each event is associated with a list of all sequence identifer (SID) and timestamps or event identifer (EID). From this format, all frequent sequences can be enumerated via simple temporal joins or intersections of the id-lists. The frequency of a pattern X in $\zeta(X)$ is obtained by computing the ratio of the number of distinct SID present in its id-list by the number of sequences.

The algorithm starts by scanning the vertical database to generate all frequent events or single-event sequences having the minimum support. The second scan of the original database allows to generate two-event sequences from an inverted database (horizontal again), then a support array of length $n * n$ is created for sequences formed by frequent events having a different EID, where n is the number of frequent one-event sequences found in the first scan. Another array is created for those sequences formed by frequent events with the same EID, the size of this array is of length $(n * (n - 1))/2$. Then all possible two-event candidate sequences in an object's timelines are checked against all input sequences, if the corresponding element occurs its respective support array is incremented and kept when greater than user-specific one.

The set of k-event sequences are obtained by performing successive joins between two sequences called generator patterns. These patterns are $(k - 1)$-patterns that share the same $(k - 2)$-prefix. It is also said that two sequences of length k belong to the same equivalent class if they share a common prefix of length $k - 1$. A class is extended to the next level by constructing candidates using either a temporal of join or an equivalent join.

Let m_1 and m_2 be two patterns sharing the prefix p with respective suffix s_1 and s_2. The considered join will depend on nature of the patterns as follow:

- When m_1 and m_2 are event patterns such as $m_1 =< ps_1 >$, $m_2 =< ps_2 >$. The generated pattern is then $< ps_1s_2 >$ and its id-list is obtained doing an equivalent join between $\zeta(m_1)$ and $\zeta(m_2)$
- When m_1 is an event pattern and m_2 a sequence pattern such as $m_1 =< ps_1 >$, $m_2 =< p \rightarrow s_2 >$. The generated pattern will be $< ps_1 \rightarrow s_2 >$ whose id-list is obtained from a temporal join between $\zeta(m_1)$ and $\zeta(m_2)$
- When m_1 and m_2 are both sequence patterns such as $m_1 =< p \rightarrow s_1 >$, $m_2 =< p \rightarrow s_2 >$ and $s_1 \neq s_2$. There are three patterns resulting of their join:
  - An event pattern $< p \rightarrow s_1s_2 >$ that is an equivalent join between $\zeta(m_1)$ and $\zeta(m_2)$
  - A sequence pattern $< p \rightarrow s_1 \rightarrow s_2 >$ that is a temporal join between $\zeta(m_1)$ and $\zeta(m_2)$

    - A sequence pattern $< p \rightarrow s_2 \rightarrow s_1 >$ that is a temporal join between $\zeta(m_2)$ and $\zeta(m_1)$

- When m_1 and m_2 are both sequence patterns and $s_1 = s_2 = s$. The only generated pattern is $< p \rightarrow s \rightarrow s >$ which is obtained by a temporal join between $\zeta(m_1)$ and $\zeta(m_2)$

An equivalent join between two patterns m_1 and m_2 with respective SID a and b, produces all occurrences of pattern m_1 and m_2 that occurred at the same time. Therefore, a tuple (a, t_1) from $\zeta(m_1)$, will match a tuple (b, t_2) from $\zeta(m_2)$ when $a = b$ and $t_1 = t_2$.

A temporal join is used to find all occurrences of a pattern m_1 preceding occurrences of another pattern m_2. Then it allows to find all tuples (a, t_1) from $\zeta(m_1)$ verifying condition $a = b$ and $t_1 < t_2$ where (b, t_2) is a tuple from $\zeta(m_2)$.

The integration of length and/or width constraints has no effect on the id-lists, thus it is done with a simple checking of width and length of the prefix.

Duration limitations like the overall duration of the discovered sequences (sliding window) requires the time difference of the first and the last events, which can be obtained by including an additional last event timestamp column to the id-list. Then, the overall duration of a sequence generated by a join can be determined by substracting the timestamps of antecedent element from that of the subsequent one and adding the result to the duration of the resulting sequence.

Minimum/maximum gap constraints between adjacent elements of a sequence destroy the class self-containment property. That is why the temporal join method needs to incorporate a timestamp verification of the difference between antecedent and subsequent elements, this difference must be above/below the user-specified limit. However, since a class is no longer self-contained, a simple self-join of the class members has to be changed for a join with the set of two-event-long sequences already generated. [28]

Chapter 3

Experimental Setup

3.1 Data Source

In order to discover knowledge from data not immediately available, it is necessary to understand what kind of information is available, where can be found, and how it will be retrieved.

Given that our aim is to mine sequential patterns in insider trading activity, attention has been placed upon the Dow Jones Industrial Average companies. The Dow (as it is commonly referred to) is a stock market index that has 30 large, publicly owned companies based in the United States as components. The Dow is also the most cited and most widely recognized of the stock market indices [29] thus the scope will be narrowed to insider trading happening within companies that are or were part of the index from January 1, 2004 up to June 29, 2012.

Since the SEC is the goverment institution regulating the markets in the United States, it is also where the most reliable data could be obtained regarding insider trading activity, as it is mandatory to disclose within two business days any statement of changes in beneficial ownership[1].

This kind of reports are Form 4 filings (See Figure 3.1), and are stored in SEC's Electronic Data Gathering, Analysis and Retrieval System (EDGAR) database. EDGAR performs automated collection, validation, indexing, acceptance, and forwarding of submissions by companies and others who are required by law to file forms, and the database is freely available to the public via web or FTP. The vast majority of documents are

[1]Every director, officer or owner of more than ten percent of a class of equity securities registered under Section 12 of the Securities Exchange Act of 1934 must file with the United States Securities and Exchange Commission a statement of ownership regarding such security.

now filed electronically, and from November 2002, the SEC required all foreign companies and foreign governments to file their documents via EDGAR. Before that time, electronic filing by foreign companies was voluntary.

SEC Form 4

FORM 4

UNITED STATES SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549

STATEMENT OF CHANGES IN BENEFICIAL OWNERSHIP

Filed pursuant to Section 16(a) of the Securities Exchange Act of 1934
or Section 30(h) of the Investment Company Act of 1940

OMB APPROVAL	
OMB Number:	3235-0287
Expires:	December 31, 2014
Estimated average burden hours per response:	0.5

☐ Check this box if no longer subject to Section 16. Form 4 or Form 5 obligations may continue. *See* Instruction 1(b).

1. Name and Address of Reporting Person*	2. Issuer Name **and** Ticker or Trading Symbol	5. Relationship of Reporting Person(s) to Issuer (Check all applicable)
GATES WILLIAM H III	MICROSOFT CORP [MSFT]	X Director; 10% Owner
(Last) (First) (Middle)	3. Date of Earliest Transaction (Month/Day/Year) 07/22/2008	Officer (give title below); Other (specify below)
ONE MICROSOFT WAY	4. If Amendment, Date of Original Filed (Month/Day/Year)	6. Individual or Joint/Group Filing (Check Applicable Line)
(Street)		X Form filed by One Reporting Person
REDMOND WA 98052		Form filed by More than One Reporting Person
(City) (State) (Zip)		

Table I - Non-Derivative Securities Acquired, Disposed of, or Beneficially Owned

1. Title of Security (Instr. 3)	2. Transaction Date (Month/Day/Year)	2A. Deemed Execution Date, if any (Month/Day/Year)	3. Transaction Code (Instr. 8)		4. Securities Acquired (A) or Disposed Of (D) (Instr. 3, 4 and 5)			5. Amount of Securities Beneficially Owned Following Reported Transaction(s) (Instr. 3 and 4)	6. Ownership Form: Direct (D) or Indirect (I) (Instr. 4)	7. Nature of Indirect Beneficial Ownership (Instr. 4)
			Code	V	Amount	(A) or (D)	Price			
Common Stock	07/22/2008		S		1,000,000	D	$25.6332 (1)	811,130,446 (2)	D	

Table II - Derivative Securities Acquired, Disposed of, or Beneficially Owned (e.g., puts, calls, warrants, options, convertible securities)

1. Title of Derivative Security (Instr. 3)	2. Conversion or Exercise Price of Derivative Security	3. Transaction Date (Month/Day/Year)	3A. Deemed Execution Date, if any (Month/Day/Year)	4. Transaction Code (Instr. 8)		5. Number of Derivative Securities Acquired (A) or Disposed of (D) (Instr. 3, 4 and 5)		6. Date Exercisable and Expiration Date (Month/Day/Year)		7. Title and Amount of Securities Underlying Derivative Security (Instr. 3 and 4)		8. Price of Derivative Security (Instr. 5)	9. Number of derivative Securities Beneficially Owned Following Reported Transaction(s) (Instr. 4)	10. Ownership Form: Direct (D) or Indirect (I) (Instr. 4)	11. Nature of Indirect Beneficial Ownership (Instr. 4)
				Code	V	(A)	(D)	Date Exercisable	Expiration Date	Title	Amount or Number of Shares				

Explanation of Responses:

1. This transaction was executed in multiple trades at prices ranging from $25.49 to $25.90. The price reported above reflects the weighted average purchase price. The reporting person hereby undertakes to provide upon request to the SEC staff, the issuer or a security holder of the issuer full information regarding the number of shares and prices at which the transactions were effected.
2. In addition, there are 424,816 shares owned by reporting person's spouse. The reporting person disclaims beneficial ownership of these securities, and this report shall not be deemed an admission that the reporting person is the beneficial owner of the securities for purposes of Section 16 or for any other purposes.

Remarks:

* Duly authorized under Special Power of Attorney appointing Michael Larson attorney-in-fact, dated February 3, 2006, by and on behalf of William H. Gates III, filed as Exhibit 99.1 to Cascade Investment, L.L.C.'s Amendment No. 2 to Schedule 13G with respect to Arch Capital Group Ltd. on March 7, 2006, SEC File No. 005-45257, and incorporated by reference herein.

William H. Gates III By: /s/ Michael Larson*, Attorney-In-Fact 07/24/2008
** Signature of Reporting Person Date

Reminder: Report on a separate line for each class of securities beneficially owned directly or indirectly.
* If the form is filed by more than one reporting person, *see* Instruction 4 (b)(v).
** Intentional misstatements or omissions of facts constitute Federal Criminal Violations *See* 18 U.S.C. 1001 and 15 U.S.C. 78ff(a).
Note: File three copies of this Form, one of which must be manually signed. If space is insufficient, *see* Instruction 6 for procedure.
Persons who respond to the collection of information contained in this form are not required to respond unless the form displays a currently valid OMB Number.

FIGURE 3.1: SEC's Form 4 filing example

3.2 Data Gathering

The gathering process required an active internet connection, plus an active R environment[2] with XML [30] and plyr [31] packages installed. In order to retrieve only the files containing SEC's form 4 filings from Dow components between 2004 and 2012, first it is necessary to filter out non-matching URLs from a quarterly master index table. To be able to select only Dow components, a mapping from text to CIK[2] was needed to be done apart given that files are associated to CIK identifiers instead of symbols or text.

Once files are downloaded, each document had to be parsed using XPath syntax to find XML nodes that match our particular criteria. Such extract criteria selected the following nodes:

- Issuer child nodes including CIK, name, and trading symbol
- Reporting owner name child node including only the first listed name if more than one owner exists
- Reporting owner relationship node including only the checked child node corresponding to either Director, Officer, Ten Percent Owner, Other, or any combination between them.
- Non-derivative[3] transaction child nodes including security title, transaction date[4], transaction size, and post-transaction number of shares owned
- The last extracted feature is the SIC[5] code, which is not located inside any of the XML tags, but in a preceding text within brackets.

Finally, if there is more than one non-derivative transaction node, then it is considered as a separate record differing only in its child nodes. In case one of the extracted child nodes is non-existent but at least one of their siblings exists, the missing ones were assigned a NA value.

3.3 Data Exploration and Pre-processing

In this part a first approximation of the retrieved collection of files containing SEC's form 4 filings is performed. The data exploration stage is about describing the data using statistical and visualization techniques. It is meant to provide important aspects of data into focus for further analysis, as well as paving the way for an optimal pre-processing stage to take place.

[2]The Central Index Key (CIK) is used on the SEC's computer systems to identify corporations and individual people who have filed disclosure with the SEC.

[3]Only open market or private purchases and sales are considered, since the quantity and timing of derivative transactions are not entirely at the discretion of the insider who receives them.

[4]If more than a single transaction was reported within the same filing, this date reflects the actual date the transaction was made.

[5]The Standard Industrial Classification (SIC) is a United States government system for classifying industries by a four-digit code.

To proceed with visualization of data features a profile of most of the nominal ones are represented by using bar charts, representing the values on an axis, and a bar of length proportional to the frequency in its perpendicular axis. For those features having a cardinal or continuous nature, histograms are used.

The retrieved raw dataset is comprised of 59200 transactions and the following 12 features: CIK, company, symbol, SIC, owner's name, owner's relation, type of stock, date, transaction size by number of shares, type of transaction (buying or selling), and remaining number of shares after transaction.

In figure 3.2, the number of total transactions by company are observed, where the most frequent ones are highlighted (See Table 3.1) and summarized by symbol.

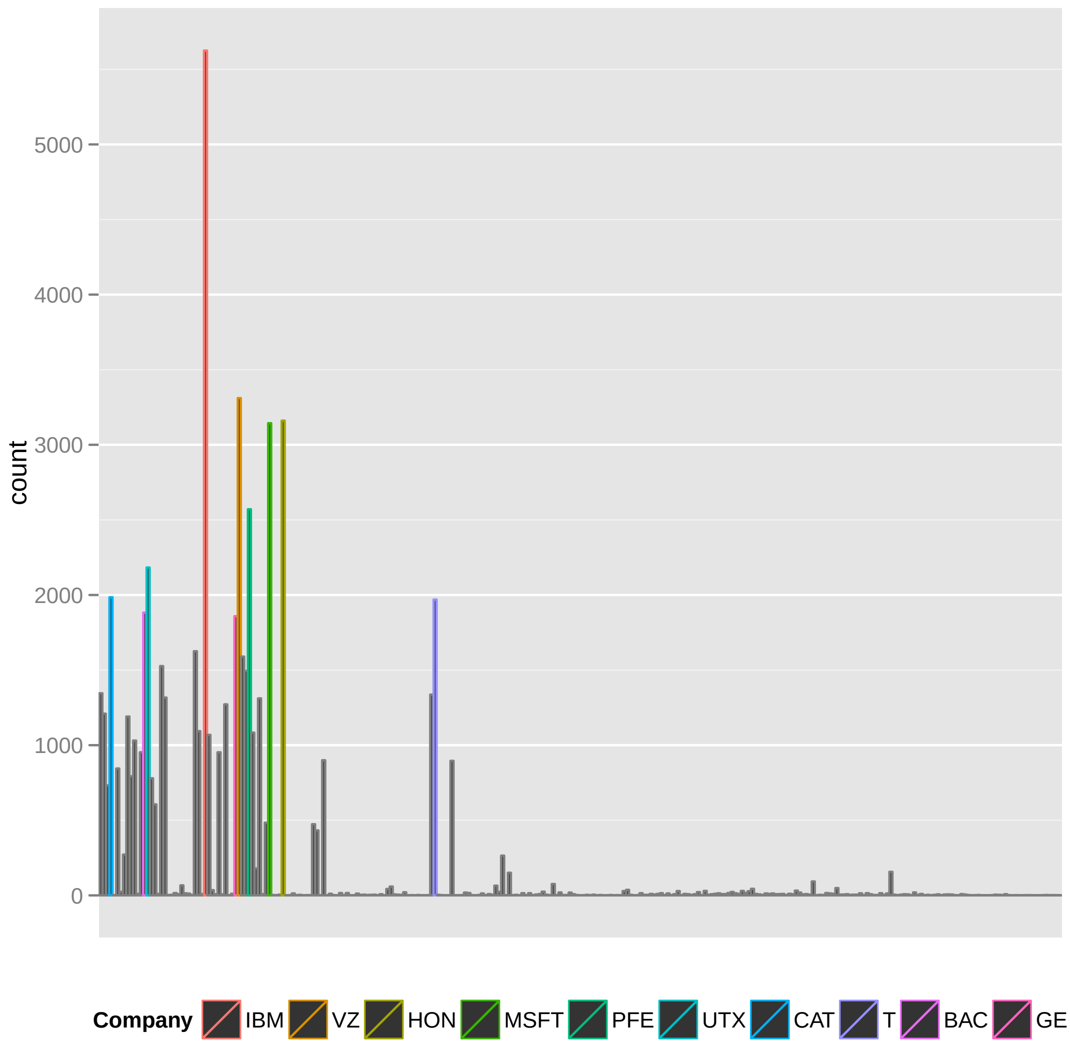

Figure 3.2: Transactions' count by Company

Frequency	CIK	Symbol	Company Name
5625	0000051143	IBM	INTERNATIONAL BUSINESS MACHINES CORP
3311	0000732712	VZ	VERIZON COMMUNICATIONS INC
3161	0000773840	HON	HONEYWELL INTERNATIONAL INC
3144	0000789019	MSFT	MICROSOFT CORP
2571	0000078003	PFE	PFIZER INC
2183	0000101829	UTX	UNITED TECHNOLOGIES CORP DE
1985	0000018230	CAT	CATERPILLAR INC
1970	0000732717	T	AT&T INC
1883	0000070858	BAC	BANK OF AMERICA CORP
1859	0000040545	GE	GENERAL ELECTRIC CO

TABLE 3.1: Top 10 Transactions' count by Company

This variable is just a less descriptive version of the company field, and it seemed to be also the same as the CIK, nonetheless the total number of items is slightly different among the three of them. The approximate total number of items is also unexpectedly greater (CIK = 263, *Company* = 275, *Symbol* = 285) than the maximum number of companies selected (35)[6]. This discrepancy was found to be due to several reasons, from companies having mutual trust funds under the same CIK, mergers which have had a renaming effect, to typos and sometimes unmatched letter cases. Specific situations where there was a need to modify a ticker symbol and/or company include the following:

- Symbol MMMC had to be changed to MMM
- Company SBC Communications Inc along with its symbol SBC were modified to the current AT&T and T company name and symbol
- Company ChevronTexaco was updated to just Chevron
- Symbol JPMC*011 was updated to JPM
- Company Merck Sharp & Dohme was changed to Merck & Co Inc
- Company St Paul Travelers Componies Inc along with its symbol STA were modified to the current Travelers Companies and TRV company name and symbol

In figure 3.3 the same bar chart is displayed after cleansing took place, plus being enhanced with the proportion of buying and selling transactions information within the bars.

[6]Please refer to Appendix A for a complete table of the actual and former Dow components for the researched period

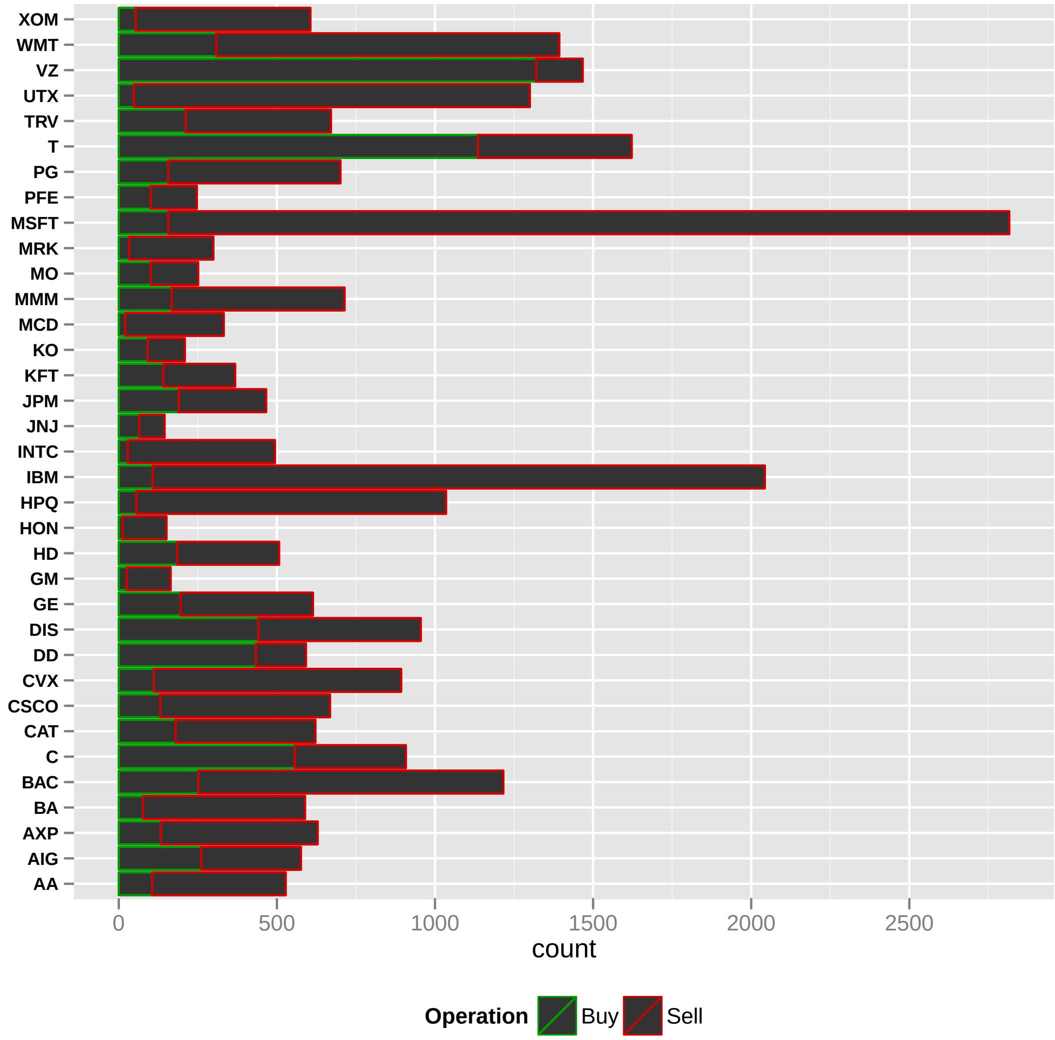

FIGURE 3.3: Filtered Transactions' count by Company

The next variable in figure 3.4 is the number of transactions summarized by industry. For its unique representation within filings, the number of transactions by SIC are pretty much the same ones before pre-processing and afterwards. The most frequent industries are also highlighted in table 3.2, where besides the SIC code the name of the industry[7] was added to improve interpretation.

A similar bar chart containing the proportion of buying and selling transactions is observed in figure 3.5. The domain of SIC codes before pre-processing was 62, and afterwards 39. The reason why even after filtering out there are still more industry codes than companies is because some of them have a business in different industries.

[7]Refer to Appendix B for a complete table of SIC/Industry pairs.

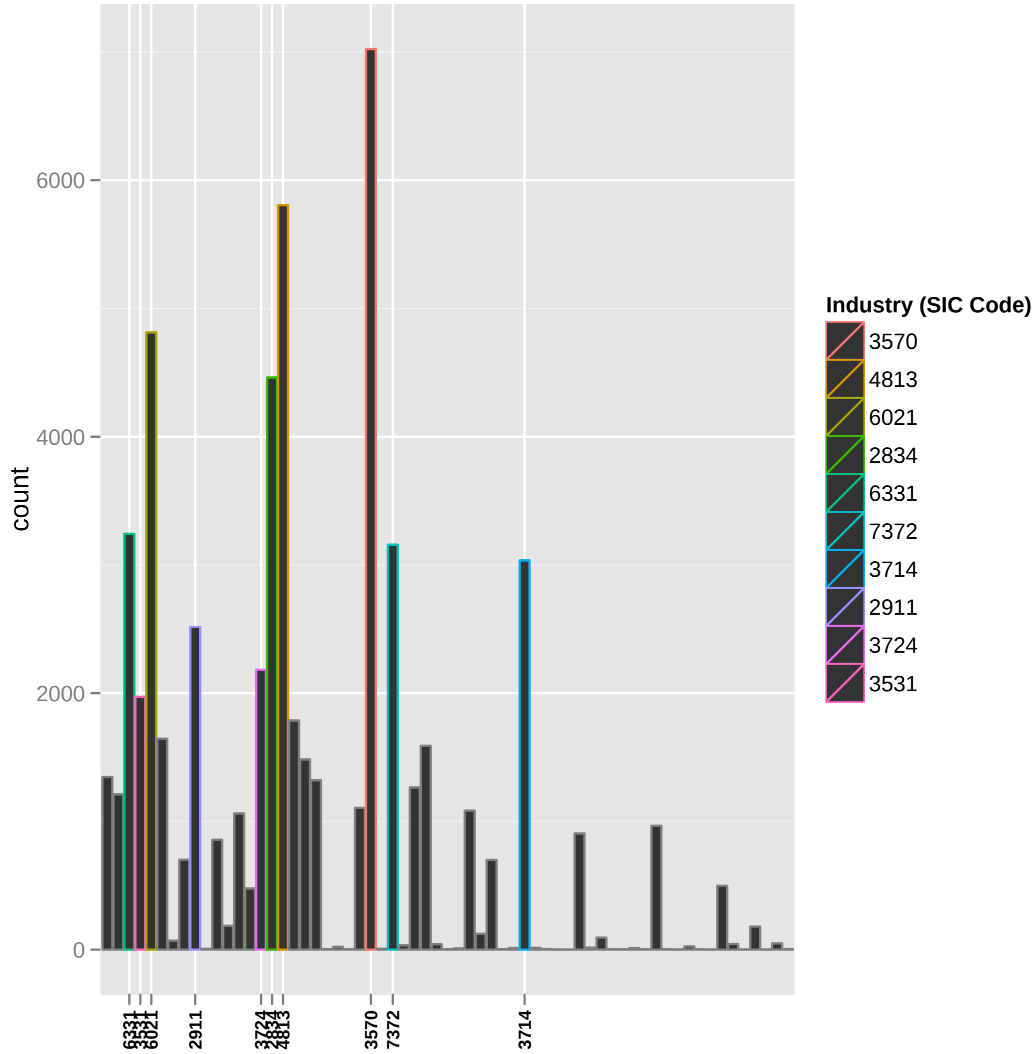

FIGURE 3.4: Transactions' count by Industry (SIC)

A third bar chart of frequencies of transactions by reporting owner is explored in figure 3.6. As it was expected, there are more owners than industries or companies, which makes it difficult to notice anything beyond the most frequent insiders. Initially there were 1615 reporting owners, and even though instances were not reduced considerably after being pre-processed, the number of unique reporting owners was consolidated at 1335. One that is hardly avoidable at first sight by the number of transactions observed, is William Gates III (Bill Gates) who filed transactions almost five times the second most active insider trader, and most of them being share disposals as observed in figure 3.7. In table 3.3 the following nine most active traders are also listed, as well as their relation and company.

Using an exploratory analysis of the first three features, it is already

Frequency	SIC	Industry
7023	3570	COMPUTER and OFFICE EQUIPMENT
5806	4813	TELEPHONE COMMUNICATIONS (NO RADIOTELEPHONE)
4813	6021	NATIONAL COMMERCIAL BANKS
4461	2834	PHARMACEUTICAL PREPARATIONS
3243	6331	FIRE, MARINE and CASUALTY INSURANCE
3158	7372	SERVICES-PREPACKAGED SOFTWARE
3035	3714	MOTOR VEHICLE PARTS & ACCESSORIES
2516	2911	PETROLEUM REFINING
2183	3724	AIRCRAFT ENGINES and ENGINE PARTS
1972	3531	CONSTRUCTION MACHINERY & EQUIP

TABLE 3.2: Top 10 Transactions' count by Industry (SIC)

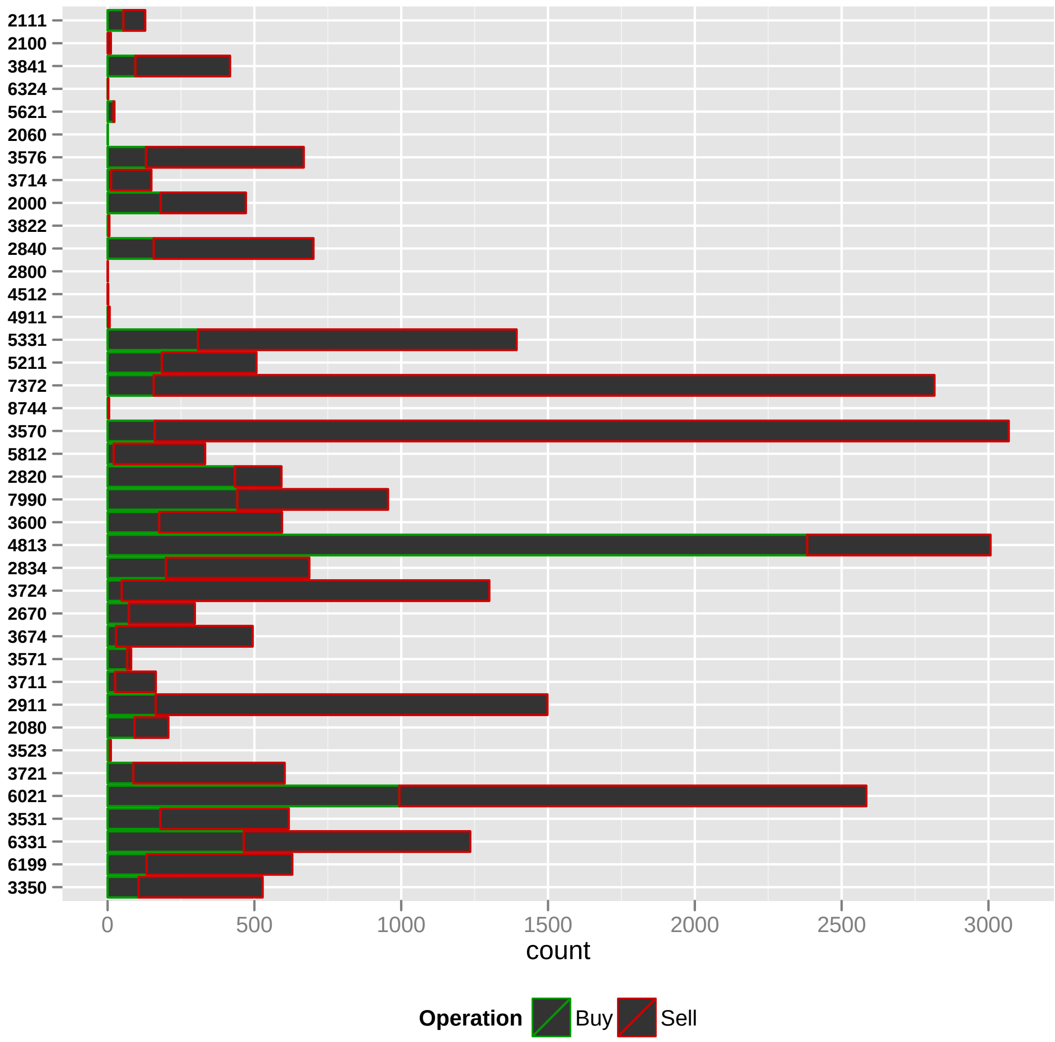

FIGURE 3.5: Filtered Transactions' count by Industry (SIC)

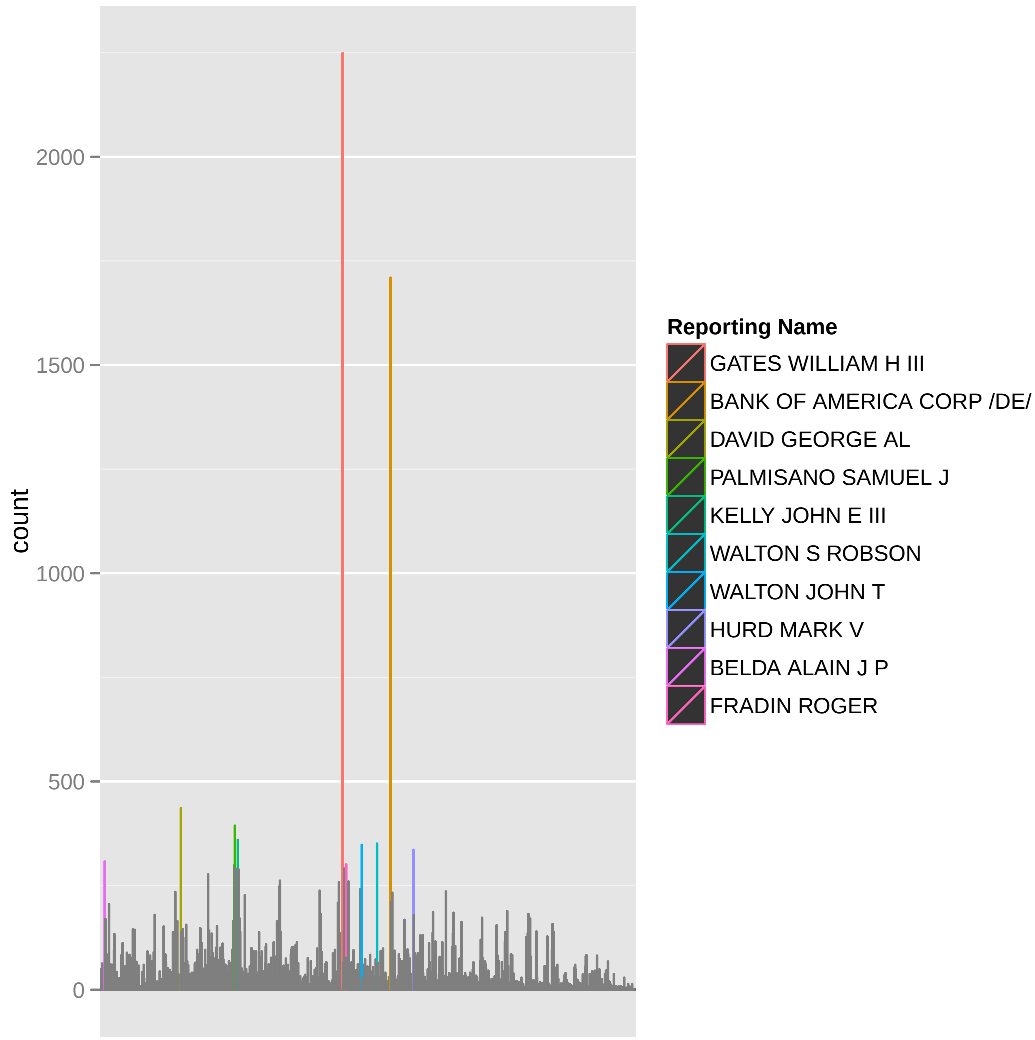

FIGURE 3.6: Transactions' count by Reporting Owner

Frequency	Name	Company
2249	GATES WILLIAM H III	MSFT
1710	BANK OF AMERICA CORP /DE/	CMGI
436	DAVID GEORGE AL	UTX
394	PALMISANO SAMUEL J	IBM
360	KELLY JOHN E III	IBM
351	WALTON S ROBSON	WMT
348	WALTON JOHN T	WMT
336	HURD MARK V	HPQ
308	BELDA ALAIN J P	AA
301	FRADIN ROGER	HON

TABLE 3.3: Top 10 Transactions' count by Reporting Owner

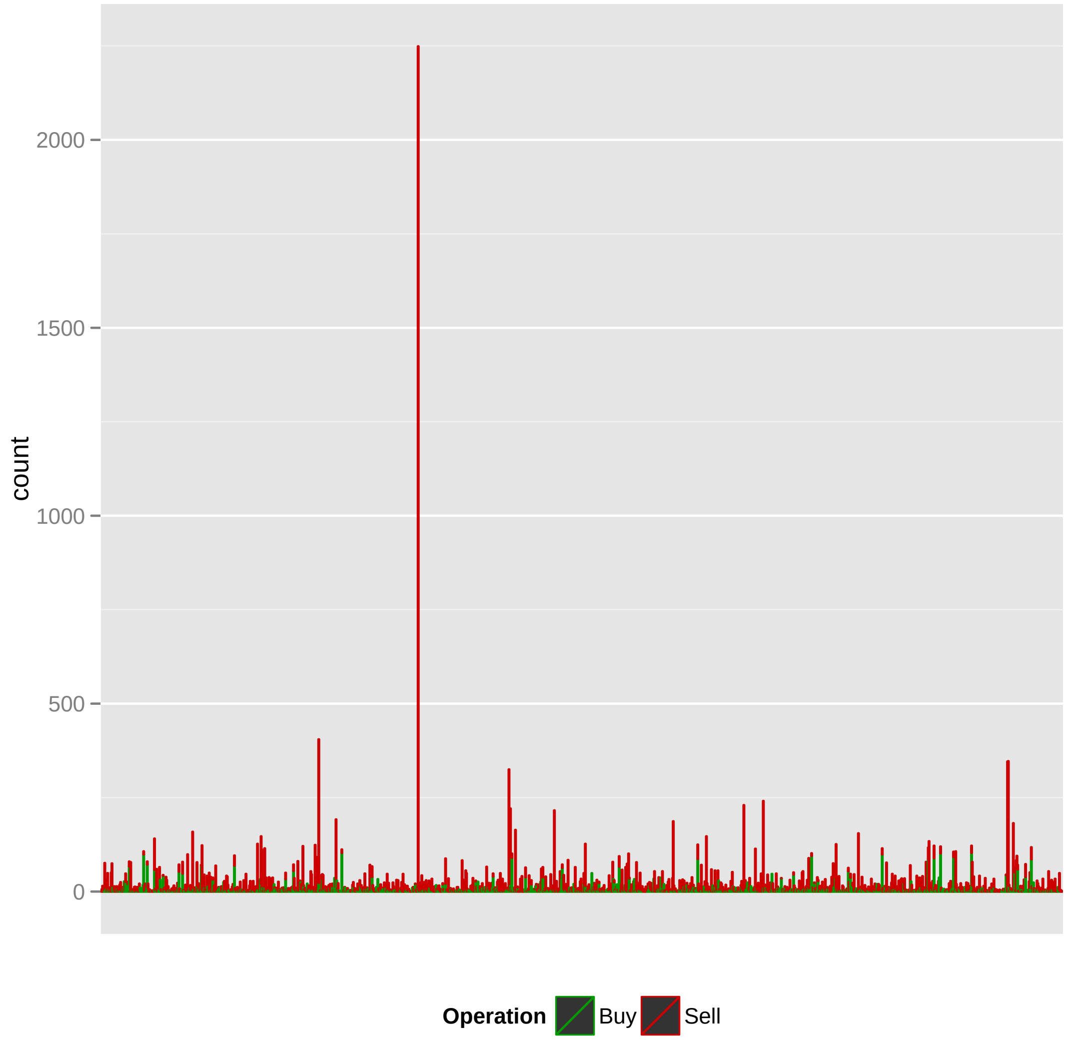

Figure 3.7: Filtered Transactions' count by Reporting Owner

highlighted the significant proportion of disposals with respect to acquisitions. It can be inferred that insiders tend to acquire the shares they own either because they created the company, are granted stocks as compensation for working there, or by derivative transactions.

In figure 3.8, the type of stock is explored only to find out there are only two significant values representing the whole feature. Common Stocks different spellings and descriptions are referring to the same security title, while the other frequent label is related to missing values. Since the scope of the analysis is focused only on open market and private transactions, transactions having missing values as type of stock are discarded. Furthermore, being left only with common stocks and its different spellings and descriptions, the whole feature is discarded because it does not provide relevant information.

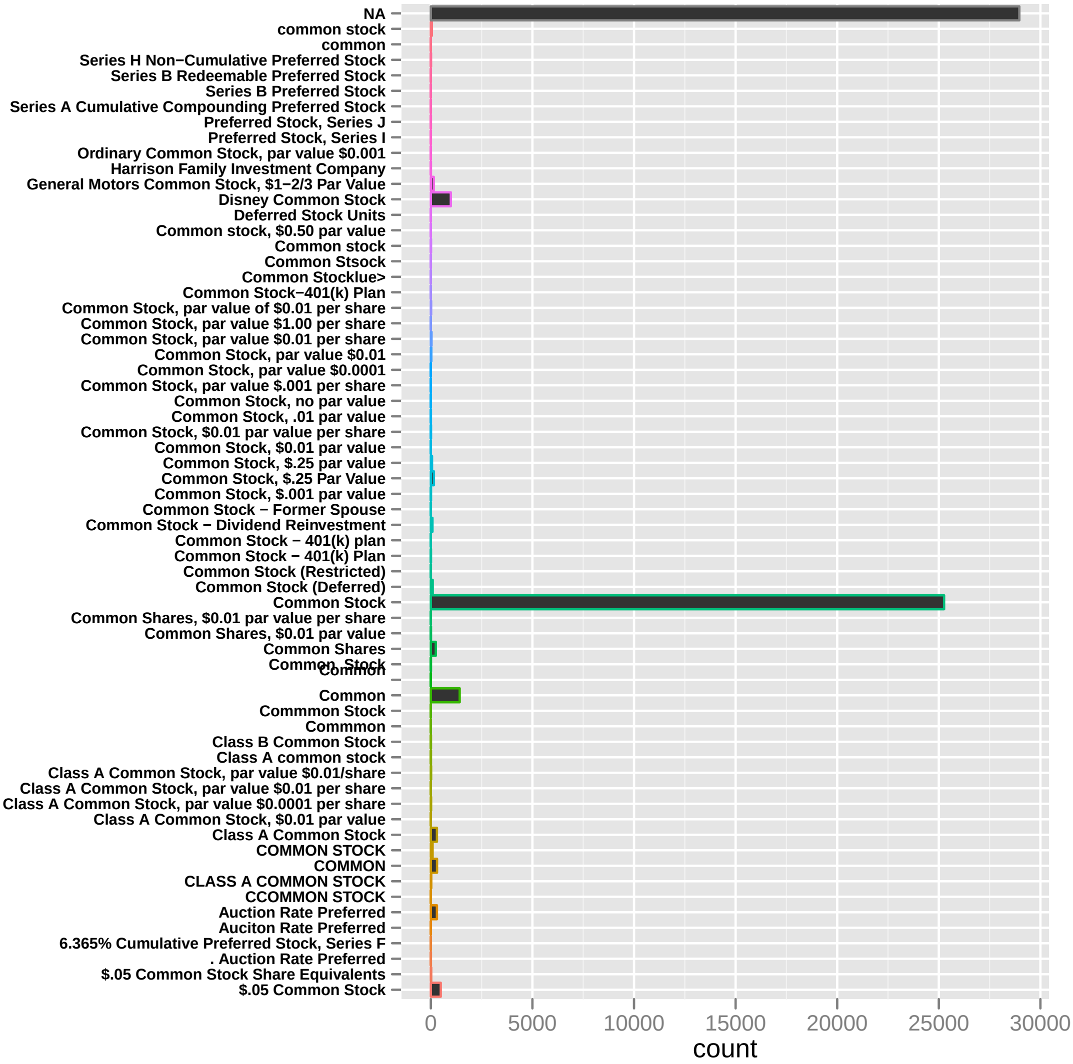

FIGURE 3.8: Transactions' count by Type of Stock

The reporting owner's relation frequency is summarized in figure 3.9. It is observable that the proportions before and after the pre-processing are somehow different at least by the count of director's transactions, which is significantly lower afterwards. Though the number of overall transactions are naturally lower given the filtering. The chart to the right representing the filtered data also displays the proportion of acquisitions/disposals for each value.

In figure 3.10, a basic comparison of frequency count by operation before (left) and after (right) filtering is measured. Once again, the missing values represent transactions where there was no open market or private acquisition/disposal operation. There are also a couple of unrecognized values in the raw dataset, whose transactions were discarded as part of the filtering out.

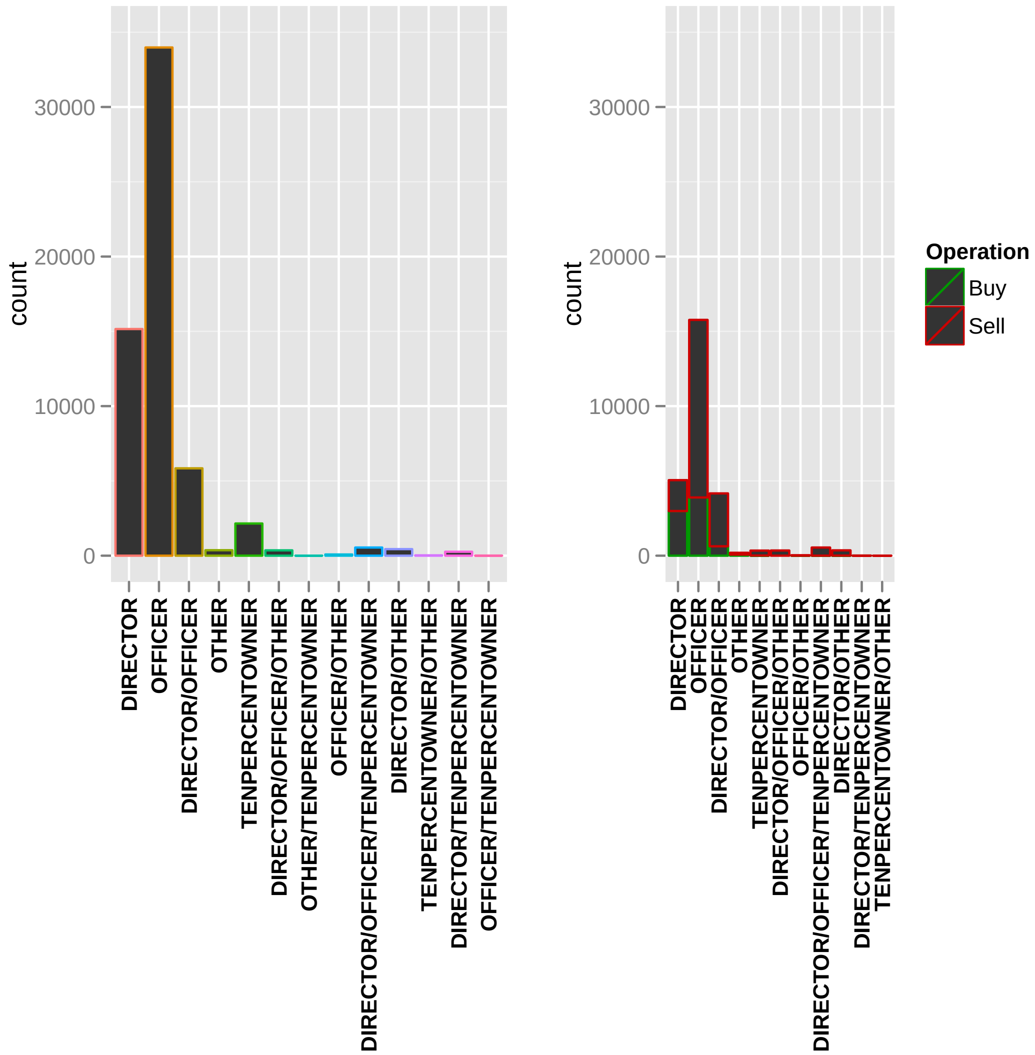

FIGURE 3.9: Comparison of Transactions' count by Owner's Relation

Figures 3.11 and 3.12 correspond to the estimated distribution of the number of transactions from January 1, 2004 to June 29, 2012. Nonetheless, it can be observed that figure 3.11 comprises also earlier and future dates. Even though no transaction was supposed to be retrieved before 2004, some have their original date before this year because an amendment is what is being retrieved. The future transaction (dated 2020-01-12) is probably a typo, or maybe a derivative expiring that year, nonetheless it is discarded as part of the cleaning stage. In figure 3.12 it is observable that many selling transactions took place around the second semester of 2007. Close to this period is the day Lehman Brothers was declared bankrupt, which intuitively might lead to think that insiders had a slightly earlier reaction to the event that is remembered as the "explosion of the bubble"

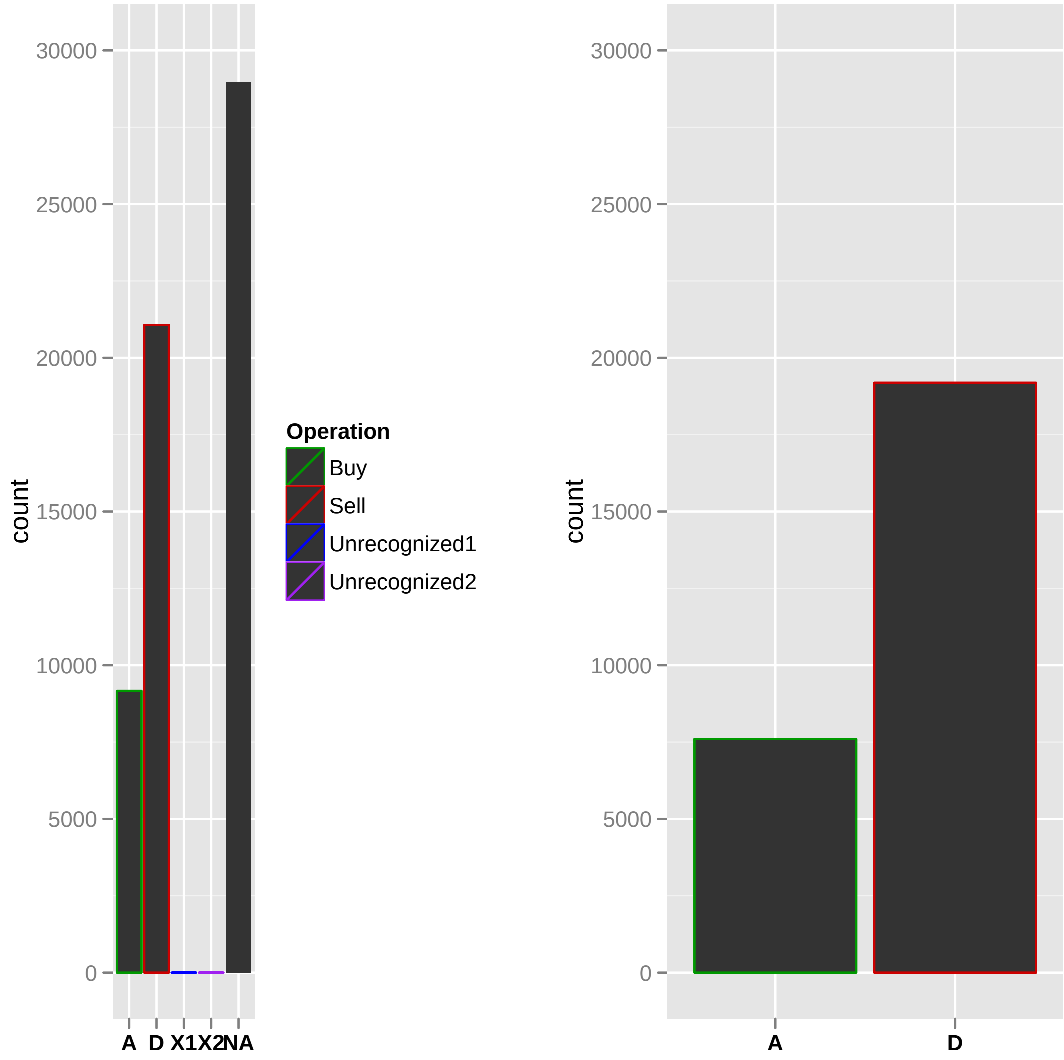

FIGURE 3.10: Comparison of Transactions' count by Operation

Figure 3.13 is a comparison between histograms of the number of traded shares per transaction, before and after pre-processing. The scale of the x axis is logarithmic, the red dotted line marks the mean, while the blue dotted line marks the median. This variable is not really useful per se given that the number of shares is a measure relative to how many shares a certain company has issued. For example, if Company *A* issued a total of 5 shares but each one of them represents a participation in the company of 20% and the company is valued at 1 million USD, then each share is worth 200 thousand USD. Figures 3.14 and 3.15 are also comparisons between histograms of the price per share, and the number of remaining shares after transaction correspondingly. Therefore, both suffer from the same lack of explanatory content by themselves. Despite of this, the price per share histogram before pre-processing allows to notice the unusual separation between the median and the mean by several orders of magnitude, thus raising concerns

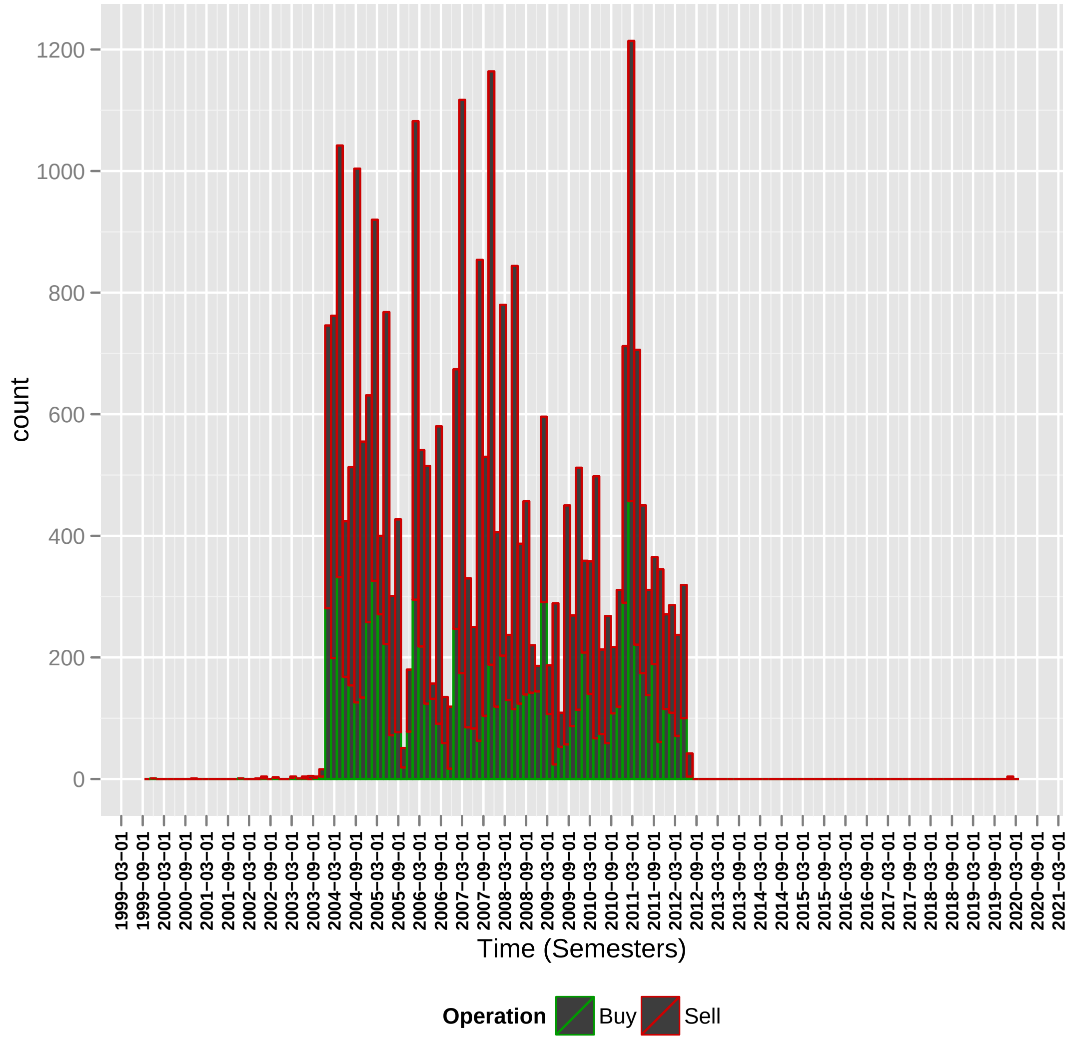

FIGURE 3.11: Histogram of Transactions' count by Date

about the validity of some of the outliers. After the transactions containing unusually high prices per share were analyzed against original records and market price at that time, it was found out that form 4 filings had typos which were corrected as part of the data cleaning stage. The fact that no share among Dow components has crossed the level of $250 USD per share served as an upper bound to pinpoint more expensive shares which immediately were known to be either typos or another type of errata.

After all the original features were analyzed, it was necessary to incorporate new features whose domain is useful for comparisons. This usefulness was not the case with features such as the number of shares, or price per share by themselves. However, a new feature like the transaction amount in USD is a valid measure of comparison between transactions. The same could be said about a feature which measures the proportion of

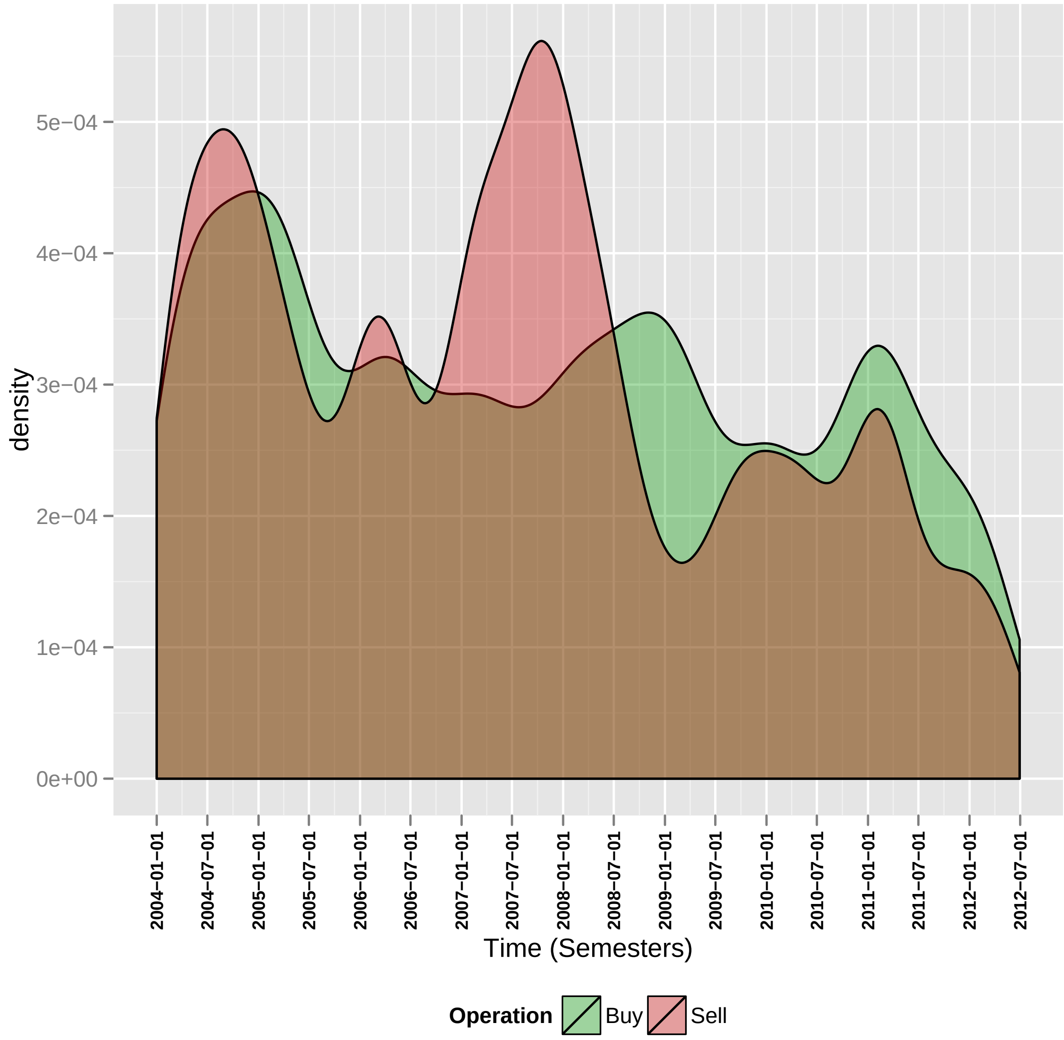

FIGURE 3.12: Filtered Density of Transactions' count by Date

shares insiders have bought/sold with respect to their own capital. For example, if two insiders sell ten of their shares, and insider #1 still has 40 shares after the operation, (s)he has sold 20% of his capital. On the other hand, if insider #2 is left with no share after selling his/her ten shares, then (s)he has taken more significant decision.

The first new feature is the transaction size in USD. It was created from the price per share, and the number of shares as the product of them. In figure 3.17 two boxplots of the same features are paired, the left one includes every outlier, while the one to the right is focused on the range of interest. The purpose of this figure as well as figure 3.16 is to provide an idea of the most relevant intervals in order to discretize the feature. The same usefulness can be exploited from figure 3.18, by looking at the distribution of both acquisitions and disposals, the scale of values where most observations take place

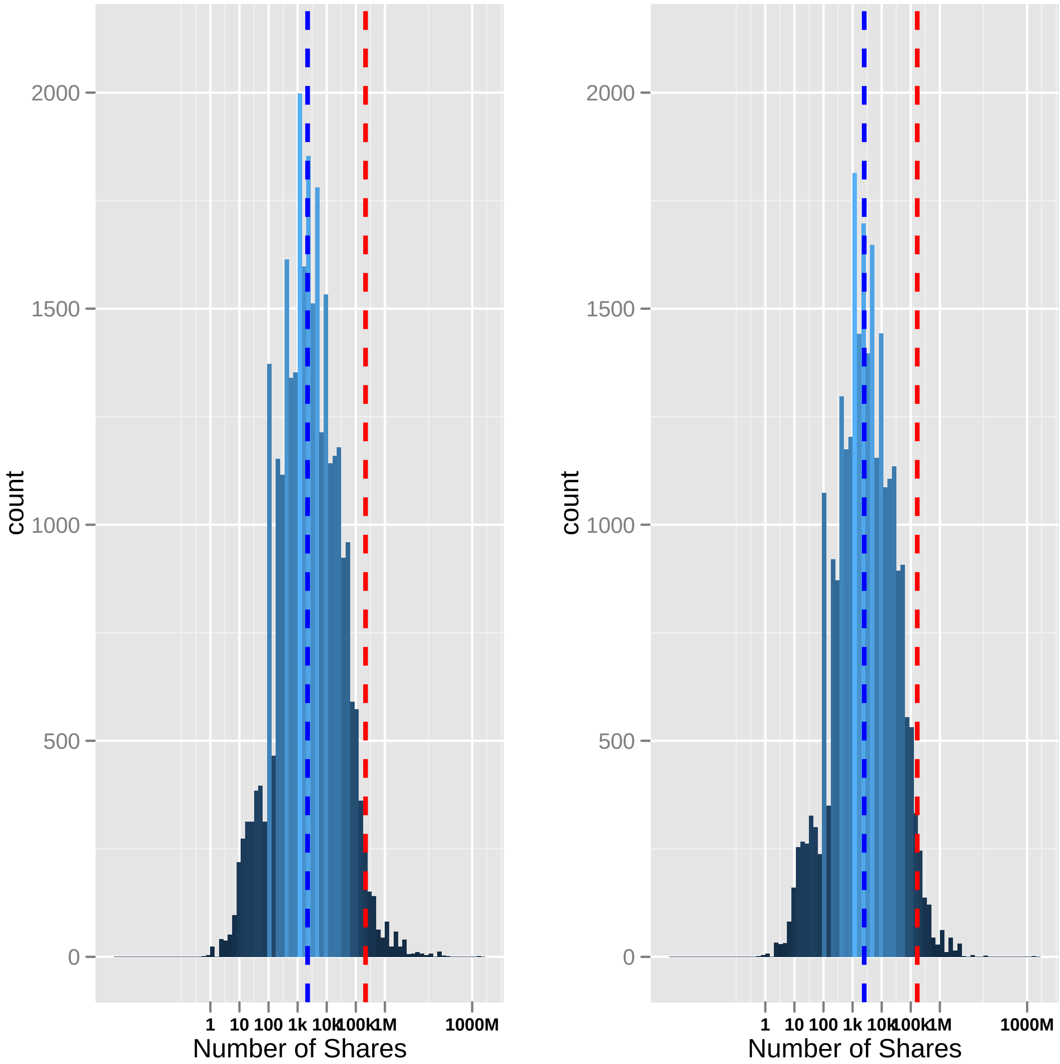

FIGURE 3.13: Histogram of Transactions' count by Number of Shares

seem to be between 1k USD and 10M USD.

Another new feature is defined as the *Ratio*. It is created to provide both information about the type of operation (buy or sell), and the magnitude of the operation relative to the quantity of shares owned beforehand. Given that we already have the information about the number of shares that were transferred in a transaction, plus the number of remaining shares after the operation, the ratio is computed as follows:

- First the number of shares before the operation is computed as the number of remaining shares minus the number of transferred shares if it was an acquisition. If it was a disposal the number of transferred shares is added to the remaining ones.

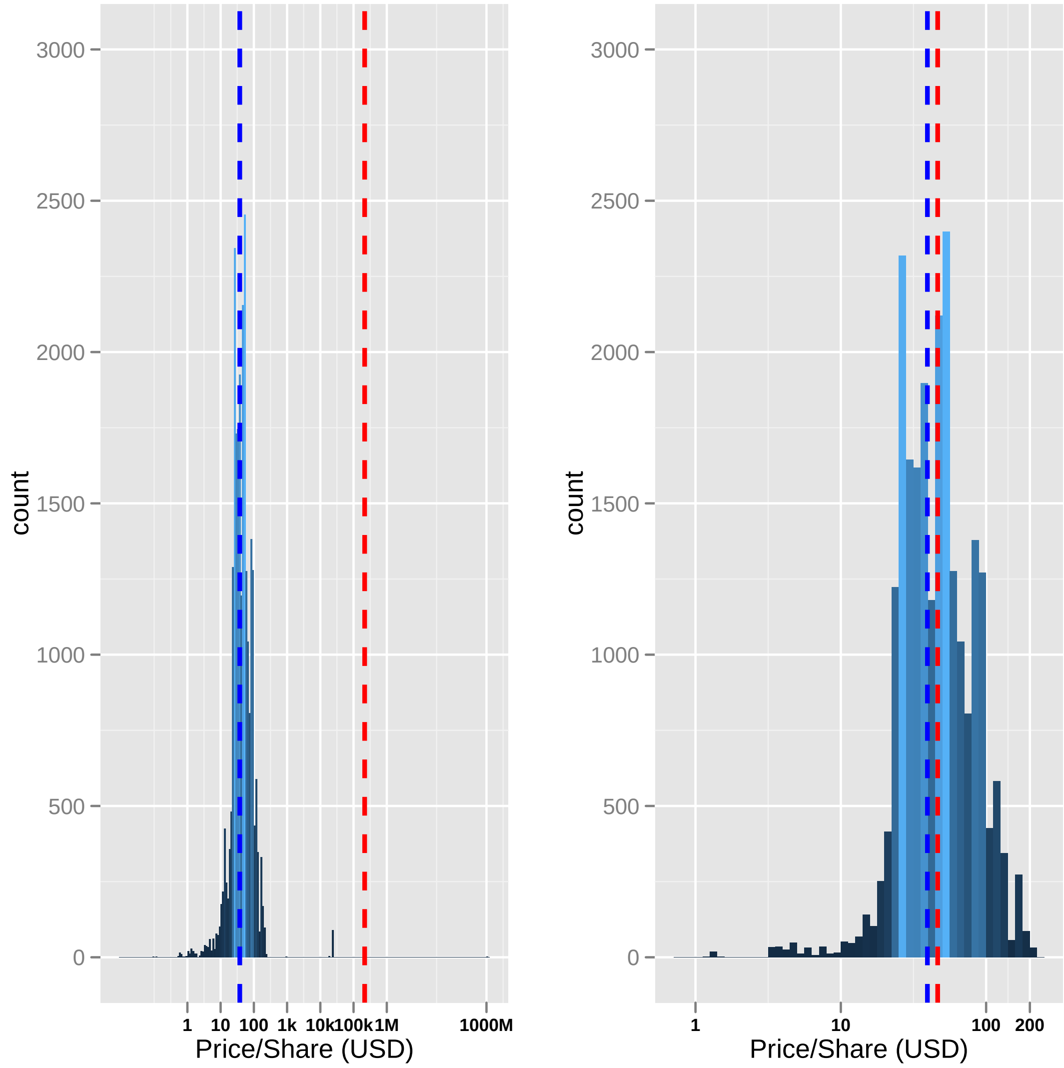

FIGURE 3.14: Histogram of Transactions' count by Price/Share

- The ratio is measured as the difference between the shares after and the ones before, divided by the shares before unless no shares were owned in the beginning, in that case the difference is divided by the number of transferred shares which will output a 1/1 or an acquisition of 100%.

The domain of this feature is between -1 and ∞, where minus one means complete disposal of owned shares, and above one would mean owning at least twice the previous stake.

The main use of figure 3.19 and 3.16 is to provide an overview of the degree of dispersion in the variable. The scatter plot being more expressive than boxplots, allows to identify thresholds more clearly than in a boxplot. Figure 3.20 is another example of how most of the observations are located around zero. Whether it is a buy or a sell operation,

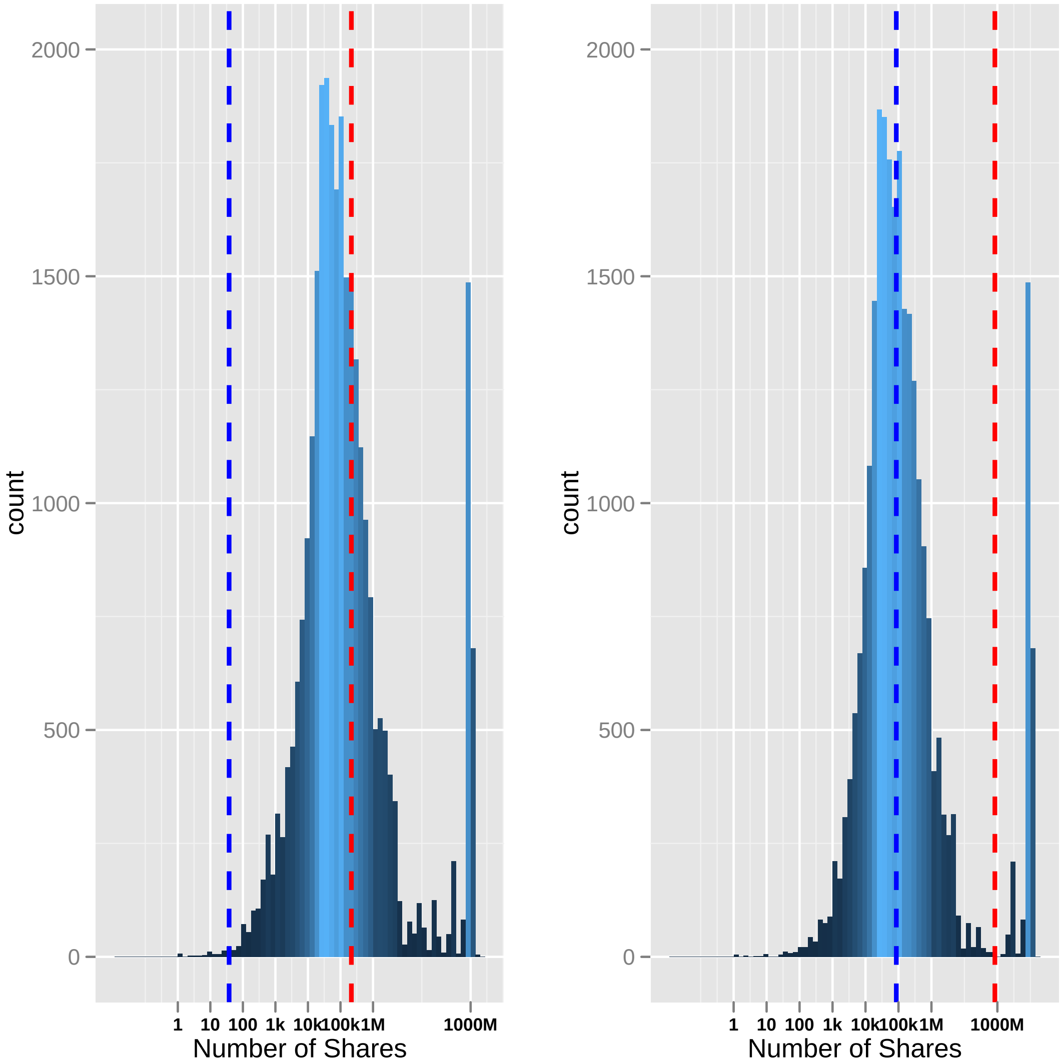

FIGURE 3.15: Histogram of Transactions' count by Remaining Number of Shares

insiders usually trade small blocks of shares they have, and seems like complete disposals and full acquisitions happen often enough to be statistically significant as well.

In order to complete the data pre-processing stage, it was needed to discretize continuous features. The number of levels depends on the domain and the amount of information to be preserved. It should be considered thta as the number of levels decreases, the average length of patterns are increased. Interval length and number was decided by analyzing the different histograms, boxplots, scatter plot figures which provided a notion of the underlying distribution. The final number of levels includes the following labels:

- Transaction Size in USD: "0", "1-10k", "10k-50k", "50k-75k", "75k-100k", "100k-125k", "125k-150k", "150k-175k", "175k-200k", "200k-250k", "250k-300k", "300k-500k", "500k-750k", "750k-1M", "1M≤".

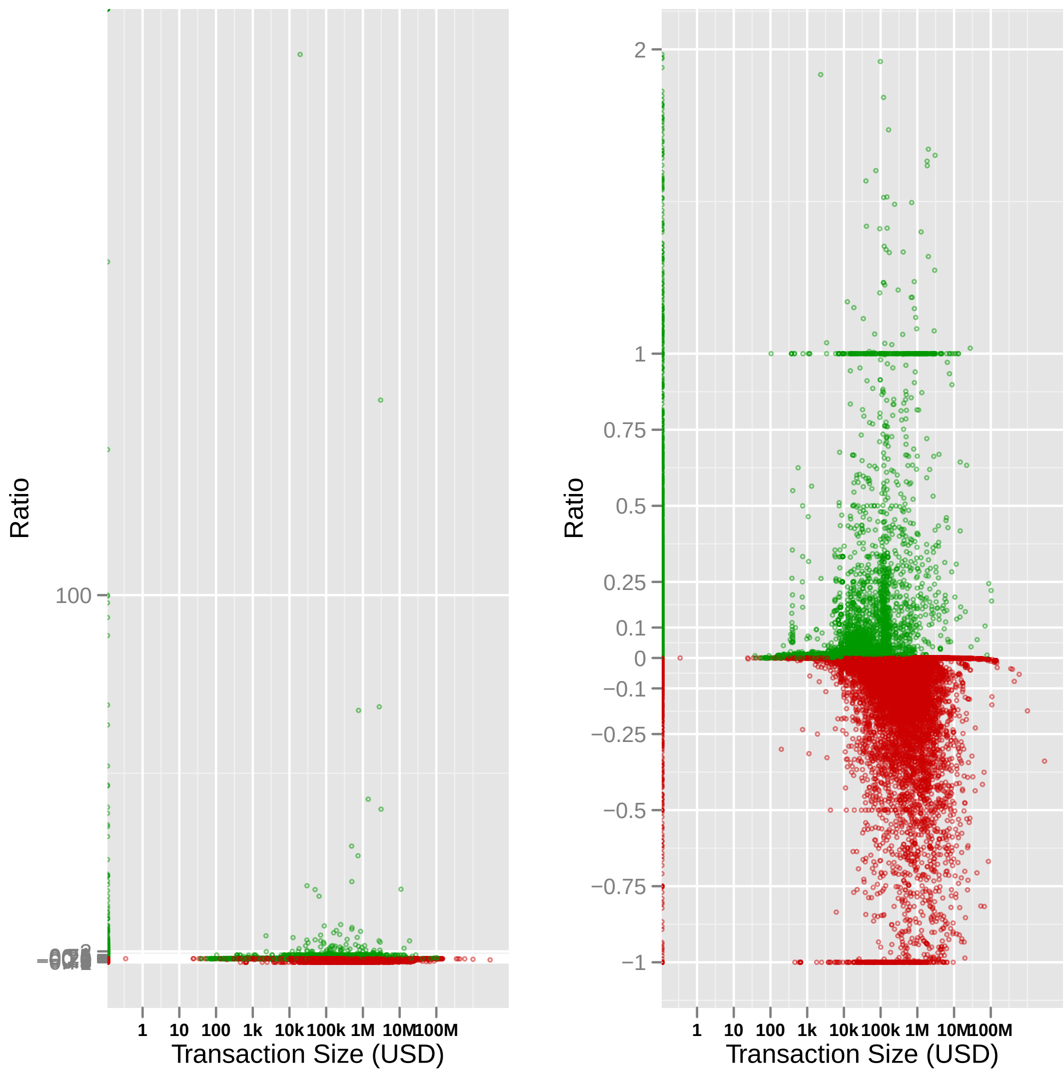

FIGURE 3.16: USD Quantity vs. Operation's Ratio Relative to Initial Shares

- Ratio: "-100%", "-[99-25)%", "-[25-10)%", "-[10-5)%", "-[5.0-2.5)%", "-[2.5-1.0)%", "-[1.0-0.5)%", "-[0.5-0)%", "[0-0.5)%", "[0.5-1.0)%", "[1.0-2.5)%", "[2.5-5.0)%", "[5-10)% ","[10-25)%", "[25-99)%", "100% ≤"

After the whole pre-processing stage was finished, the base dataset had 26791 records and 14 features.

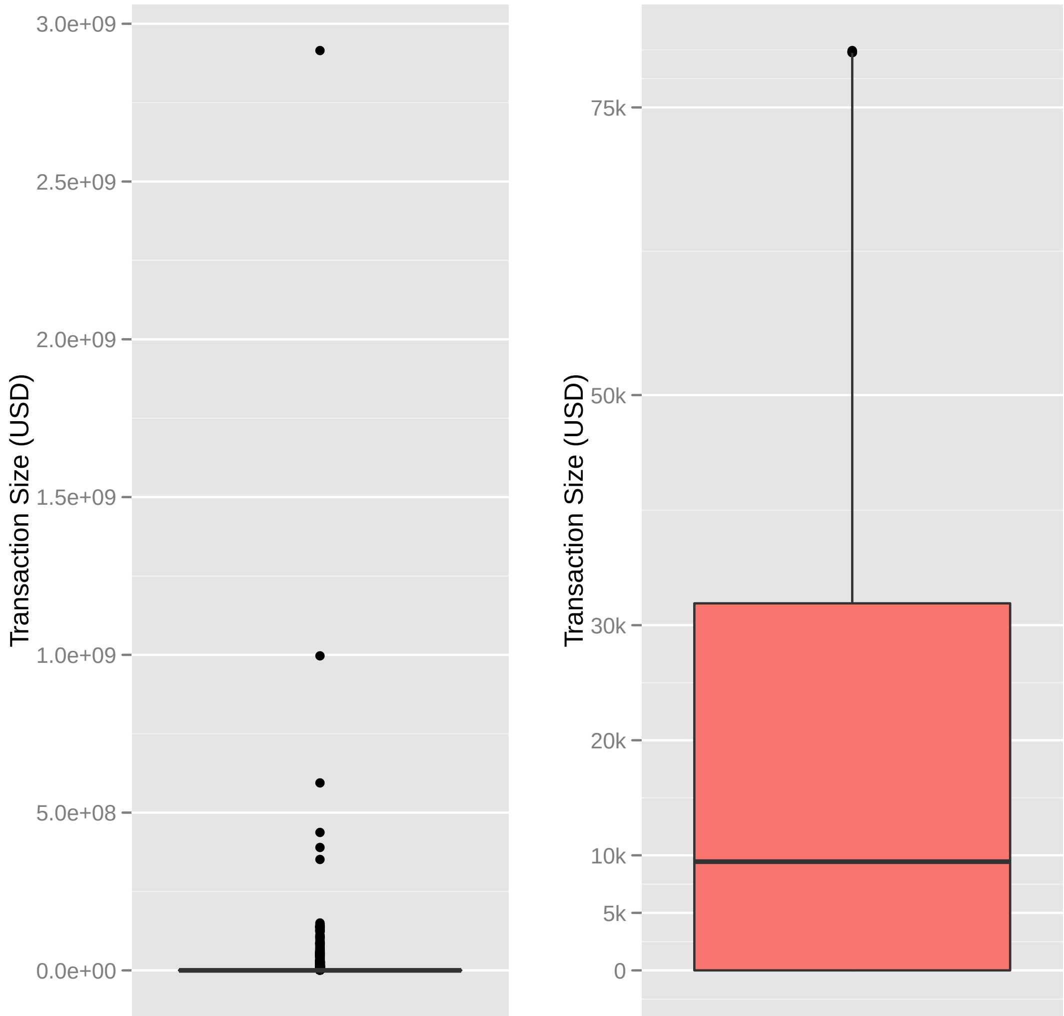

FIGURE 3.17: Boxplot USD Quantity

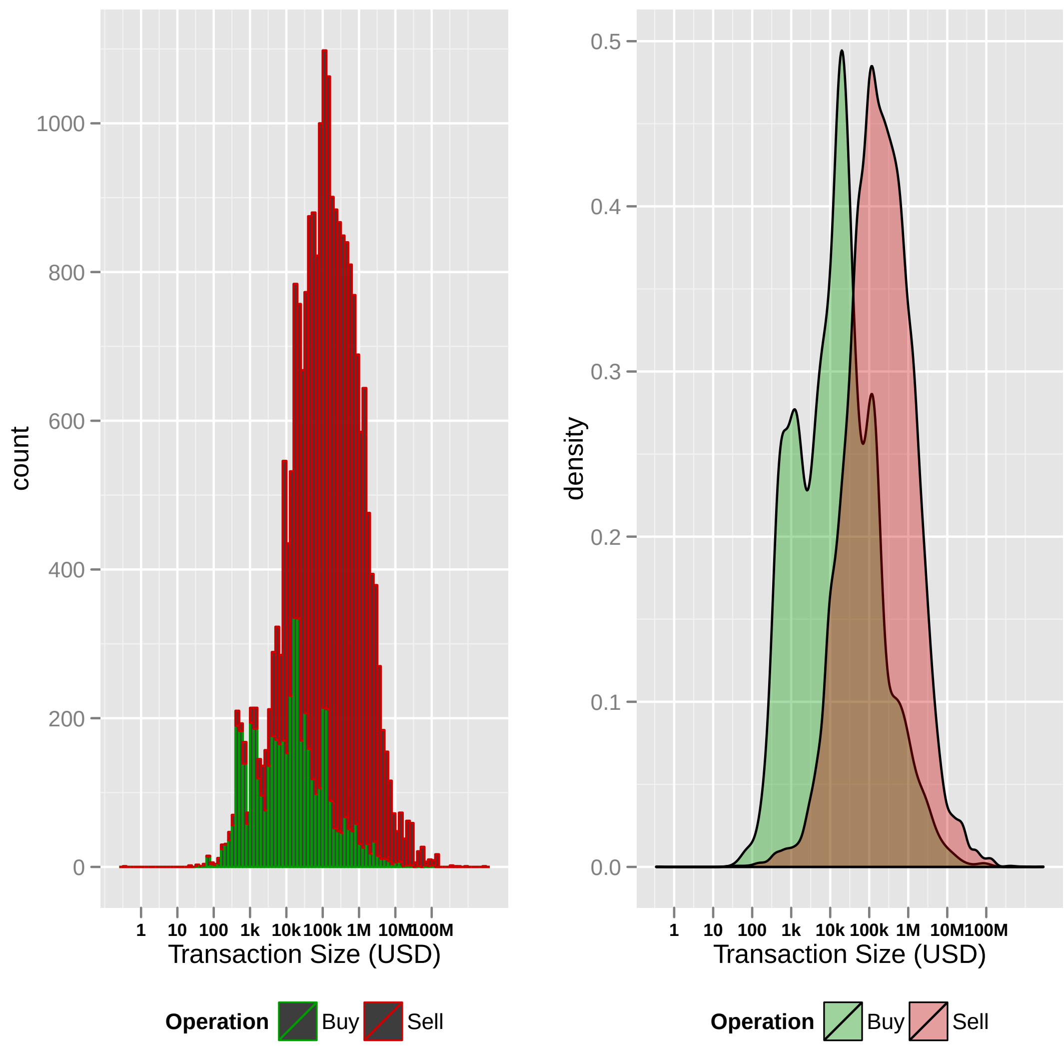

FIGURE 3.18: Filtered Transactions' count by USD Quantity

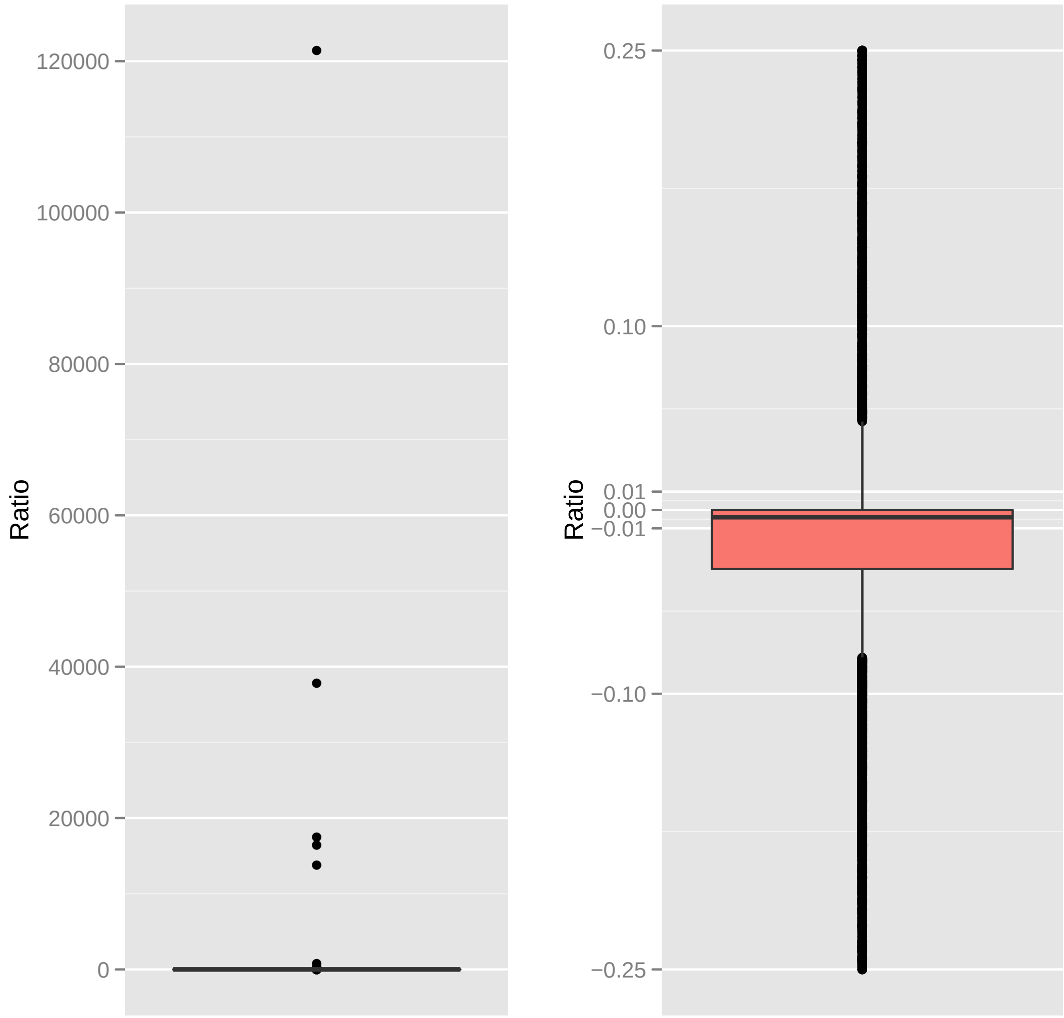

Figure 3.19: Boxplot Operation's Ratio Relative to Initial Shares

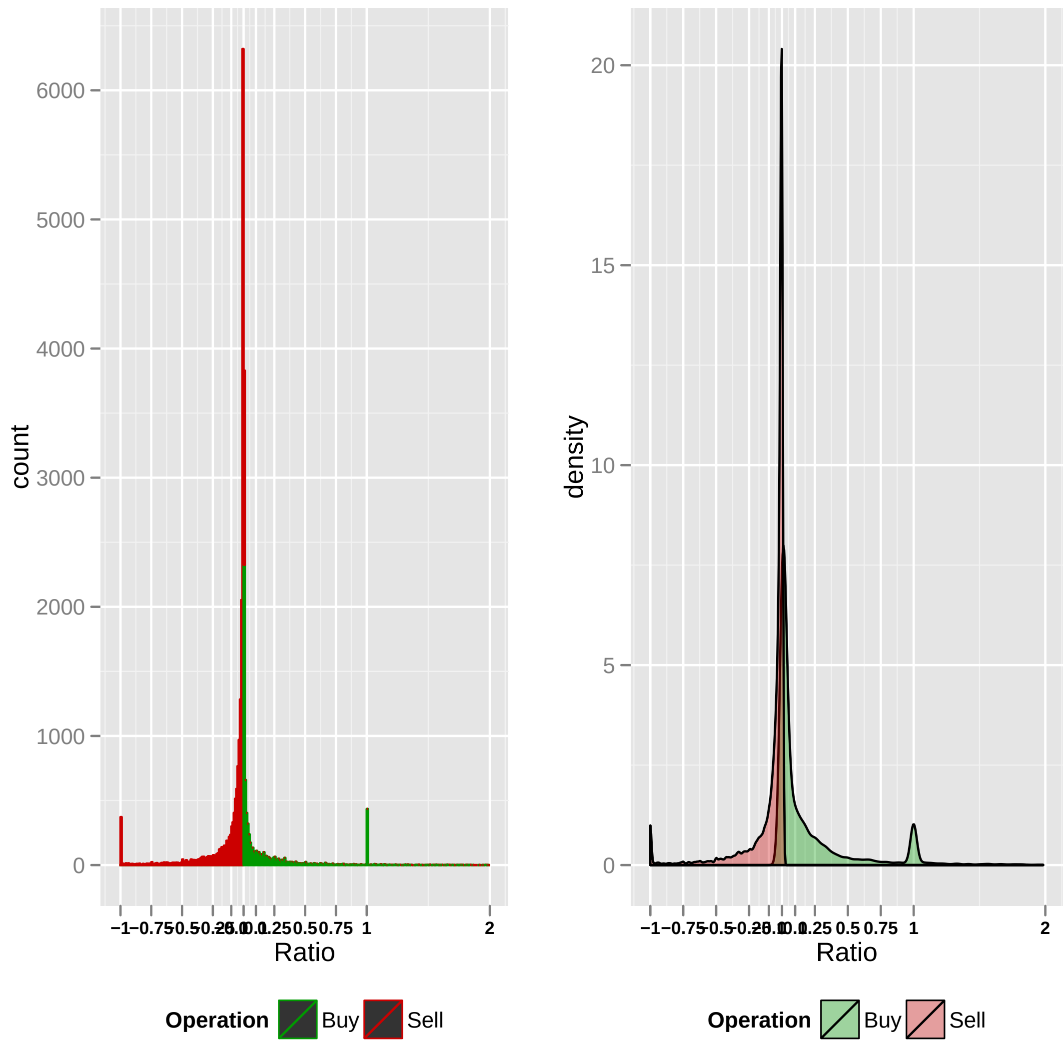

FIGURE 3.20: Filtered Transactions' count by Operation's Ratio Relative to Initial Shares

Chapter 4

Experiments

The aim of mining a sequence database is to find interesting patterns that augment the descriptive power of the raw data and that are useful to build accurate models.

The dataset was formatted to be suitable for sequence mining, it is composed by relevant data concerning open market or private operations comprising the last eight years. Two contexts are derived from our dataset, the breakpoint is September 15, 2008, the day Lehman Brothers went bankrupt, also considered to be a turning point at the beginning of the latest financial crisis.

These two contexts represent two sets of characteristic situations which can be considered as classes or labels, and from them, correspondent sequence rules are mined to compare how different is one set of rules from another. Furthermore, additional segmentation of input sequences are evaluated, represented as the following types of input sequences or SIDs:

- Common sequences by Industry
- Common sequences by Company
- Common sequences by Individual

Each one of them is used to to find out if there are interesting sequences of events common to its segment. Despite the fact several patterns might be common in all three segments, the idea is to validate either those recurrent similarities or rule out the possibility of common patterns at this level. Moreover, a financial expert might use the information provided in the comparisons, and be able to identify potential signals specific to a trend.

A second comparison stage takes place at a context level (See table 4.1). Those patterns associated to some minimum quality measures but also discriminant with respect to the context, are the most significant ones, and gathering them is also considered to be the main objective in the experiments.

Several approaches were tried from the beginning of the experiments, initial issues were regarding how our features had to be coerced to resemble the market basket paradigm exploited by the cSPADE algorithm.

Initially, we let each feature be an item's category, and therefore each attribute's value was an item by itself. The main problem with this item representation was the inherent lack of connection between items. Such items could have been frequent, but contained no meaningful information as sequences or rules. This lack of connection gave place to rules like: *"If A, B are observed then C follows"*
Where A could be the insider's relationship, and B the transaction amount. Due to this lack of connection, A could have its support in a set of sequences, but it meant B's support may not necessarily be the same one. The intended rule should have been: *"When A trades B amount within a certain time period, it follows an event C"*
However, what was achieved resembled a rule such as: *"When A trades some unknown amount, and B amount is traded by unknown insiders within a certain time period, it follows an event C"*
These kinds of rules did not make sense for our purposes.

To overcome this problem, items had to be combinations of at least two attributes. The next approach consisted of adding pair-wise relationships between some attributes, as new item categories. After experiments were done using this representation, it was found out that the obtained sequences and rules contained more events than they should, and those events were mostly redundant. An attribute's value could have been twice in the same event, first time as a single item, and the second time inside one of the newly created pair-wise items.

Simultaneously, another restriction had to be taken into account. Input sequences identified by a sequence ID (SID), enforced an exclusion of attributes depending on what kind of SID was being used. For example, if SIC code was used as SID, this meant each transaction related to a specific industry was part of the same sequence and potentially no other input sequence. Therefore, a certain item from the company's attribute could be infrequent if it is only related to one industry. The same thing could happen in a greater or lesser degree if CIK or $InsiderID$ were used as SID.

The final approach deals with both problems, however there is a trade-off involved. The problem related to the lack of connection between items was approached again by adding relationships between items. This time instead of creating couples of items, a single attribute having as value the combination of all attributes was created, and each unique combination became part of an alphabet. The obvious trade-off is the exponential explosion of possible combinations of items. In order to limit the number possible combinations, and also avoid having items which are strictly related to the choice of SID, no matter which SID is chosen a combination of items is created from the ones that contribute with quantitative measures, plus a feature which is common among any kind of sequence input, such attributes are: Insider Relation, Transaction Amount in USD, Proportion of Shares Traded.

The experiments described in the following sections of this chapter share some common parameters in order to be comparable. However, due to the overwhelming number of possible experiments, only a reduced subset is presented given a set of parameters allowing a fair comparison.

The final experimental framework to mine sequential patterns will be using 9 different evaluation experiments per dataset. The alphabet comprised 842 items[1], where each item is actually representing three features. Constraints on the length were fixed to a maximum of 10 items per event, and 2 events per sequence. There are three different scenarios induced by time constraints, all of them restrict support count only to those frequent sequences of events separated by either a month, quarter, or a year. Pattern generation was done considering a minimum support threshold of 10%, using lower values proved to be too demanding in some of the setups given the available resources (less than 2 GB of memory). The only set of experiments using a different minimum support threshold (1%) is the one using $InsiderID$ as input sequences.

Sequence association rules were generated using a minimum confidence threshold fixed to 60%, further selection is done by considering those rules whose $lift$[2] is greater than one to be considered significant, also signaling that the probability of occurrence of the antecedent and that of the consequent are not independent of one another. The final comparison between results from both datasets is done by coercing significant rules back to sequences in order to compute the similarities between them. Jaccard similarity measure is used for this purpose, where similar sequences are quantified as the sum of the lengths of the sequences minus the length of the longest common subsequence threshold at 0.5.

[1]Please refer to Appendix C for a complete list.

[2]Lift is found by dividing the confidence by the unconditional probability of the consequent, or by dividing the support by the probability of the antecedent times the probability of the consequent.

	Pre-crisis Dataset	**Post-crisis Dataset**
Number of Items	842	842
Number of Transactions	11742	6256
Density	0.0717083972	0.1345907928

Table 4.1: Summary Dataset

4.1 Industry Patterns

Initially, industry patterns seem to be the most general among the three different types of input sequences. There are 36 industries identifying their correspondent input sequences in each dataset. The statistics about the results of the patterns extraction and rules generation are summarized in tables 4.2 and 4.3 respectively. The number of patterns generated from pre-crisis dataset is about ten times the one from post-crisis dataset no matter which kind of time constraints are taken. Maximum sequence length is also higher, but then the actual number of rules and the significant ones are almost the same between both scenarios. However, the most interesting outcome is the lack of significant rules in quarterly results for a dataset apparently having more potential rules. It can be inferred that there used to be a slightly higher tendency to trade more frequently, and possibly those habits are different lately leaving a longer gap between transactions.

	Time Constraints (days)		
	Monthly	**Quarterly**	**Yearly**
	$mingap = 28$ $maxgap = 31$ $window = 31$	$mingap = 89$ $maxgap = 92$ $window = 92$	$mingap = 364$ $maxgap = 366$ $window = 366$
Number of Patterns	12994	12753	12784
Max Sequence Length	10	10	10
Number of Rules	17	2	0
Significant Rules ($lift > 1$)	13	0	0

Table 4.2: Pre-crisis Industry Results

	Time Constraints (days)		
	Monthly	**Quarterly**	**Yearly**
	$mingap = 28$ $maxgap = 31$ $window = 31$	$mingap = 89$ $maxgap = 92$ $window = 92$	$mingap = 364$ $maxgap = 366$ $window = 366$
Number of Patterns	1924	1820	1735
Max Sequence Length	6	6	6
Number of Rules	11	7	0
Significant Rules ($lift > 1$)	11	5	0

Table 4.3: Post-crisis Industry Results

In tables 4.4 and 4.5 more detailed information of the obtained rules is summarized for each context. Given that there was only one timeframe where both datasets yielded significant rules, these are the base for a head-to-head comparison of patterns. The similarity between the two sets of rules is computed using the Jaccard similarity measure, as expected there is no similarities between both of them. Despite of their differences, what is common among the rules is the presence of a frequent item in several rules, and sometimes as both antecedent and consequent. For example, that is the case in pre-crisis dataset with item "I21" which its actual information is "DIRECTOR/OFFICER; 10k-50k; -[0.5,0]

	Quality Measures		
Rule	**Support**	**Confidence**	**Lift**
<I28,I17> → <I17>	0.1111	0.8	2.6182
<I19,I21,I28> → <I21>	0.1111	0.8	2.2154
<I19,I21,I16> → <I21>	0.1111	0.8	2.2154
<I19,I16> → <I21>	0.1111	0.8	2.2154
<I21,I16> → <I21>	0.1389	0.7143	1.978
<I19,I28> → <I21>	0.1111	0.6667	1.8462
<I20,I21,I16> → <I21>	0.1111	0.6667	1.8462
<I20,I16> → <I21>	0.1111	0.6667	1.8462
<I49,I70,I160> → <I49>	0.1111	0.8	1.2522
<I38,I60,I95> → <I95>	0.1111	0.8	1.152

Table 4.4: Pre-crisis Industry Rules(Monthly)

	Quality Measures		
Rule	**Support**	**Confidence**	**Lift**
<I62> → <I62>	0.0882	0.6	2.6182
<I6,I70,I76> → <I48>	0.0882	1	2.2154
<I76,I86> → <I70>	0.1176	1	2.2154
<I72,I76,I86> → <I70>	0.0882	1	2.2154
<I72,I76> → <I70>	0.0882	1	1.978
<I70,I76> → <I48>	0.0882	0.75	1.8462
<I101,I102,I95> → <I34>	0.0882	0.6	1.8462
<I72,I86> → <I70>	0.0882	0.75	1.8462
<I2,I76,I86> → <I70>	0.0882	0.75	1.2522
<I2,I76> → <I70>	0.0882	0.75	1.152

Table 4.5: Post-crisis Industry Rules (Monthly)

After labels are mapped back to their original meaning, the ninth rule from the pre-crisis dataset could be interpreted as follows:

- *If I49, I170, and I160 take place the same day, then after one month it is likely that I49 will happen"*

4.2 Company Patterns

Just like industry patterns were summarized, in tables 4.6 and 4.7 the main information about the mined patterns and rules is listed there. Despite being quite similar beforehand, company-based input sequences do not yield suitable significant rules to make a comparison. Compared to industry based patterns, it is expected that no yearly patterns are generated, but for the remaining scenarios the number of rules is reduced significantly. Post-crisis dataset went from having monthly significant patterns to no patterns at all. This happened because the candidate sequences are not above the minimum confidence threshold, which was expected to happen for yearly or even quarterly ones, but not for monthly sequences. Because of the absence of significant rules, no similarity comparison between rules was possible.

	Time Constraints (days)		
	Monthly	**Quarterly**	**Yearly**
	$mingap = 28$ $maxgap = 31$ $window = 31$	$mingap = 89$ $maxgap = 92$ $window = 92$	$mingap = 364$ $maxgap = 366$ $window = 366$
Number of Patterns	14101	13811	13826
Max Sequence Length	10	10	10
Number of Rules	12	1	0
Significant Rules ($lift > 1$)	7	0	0

Table 4.6: Pre-crisis Company Results

	Time Constraints (days)		
	Monthly	**Quarterly**	**Yearly**
	$mingap = 28$ $maxgap = 31$ $window = 31$	$mingap = 89$ $maxgap = 92$ $window = 92$	$mingap = 364$ $maxgap = 366$ $window = 366$
Number of Patterns	804	770	760
Max Sequence Length	5	5	5
Number of Rules	0	0	0
Significant Rules ($lift > 1$)	0	0	0

Table 4.7: Post-crisis Company Results

4.3 Individual Patterns

In tables 4.8 and 4.9 the mining results are summarized one more time. The singularity of this scenario under the specified parameters lies in the output. Neither pre-crisis scenario nor post-crisis one produced any rules. Even though minimum support was relaxed to 1% instead of 10% as in previous experiments. In the end, there were not monthly, quarterly, or yearly significant rules at all to be able to measure their similarity between scenarios.

	Time Constraints (days)		
	Monthly	**Quarterly**	**Yearly**
	$mingap = 28$ $maxgap = 31$ $window = 31$	$mingap = 89$ $maxgap = 92$ $window = 92$	$mingap = 364$ $maxgap = 366$ $window = 366$
Number of Patterns	436	435	435
Max Sequence Length	4	4	4
Number of Rules	0	0	0
Significant Rules ($lift > 1$)	0	0	0

Table 4.8: Pre-crisis Individual Results

	Time Constraints (days)		
	Monthly	**Quarterly**	**Yearly**
	$mingap = 28$ $maxgap = 31$ $window = 31$	$mingap = 89$ $maxgap = 92$ $window = 92$	$mingap = 364$ $maxgap = 366$ $window = 366$
Number of Patterns	160	160	159
Max Sequence Length	2	2	2
Number of Rules	0	0	0
Significant Rules ($lift > 1$)	0	0	0

Table 4.9: Post-crisis Individual Results

Chapter 5

Conclusions and Future Work

5.1 Conclusions

This work presented a methodology for mining temporal associations among frequent patterns occurring in insider trading filings data. A first stage consisted of pre-processing the data retrieved from SEC's EDGAR database, an extensive but necessary process to obtain a dataset suitable for a data mining project. The sequences rule generation demanded high quality measures in order to extract significant rules from the mining results. The only and possible the best way of generating such rules in both pre-crisis and post-crisis datasets was by using industry's identifier as input sequences. The resulting patterns proved to be different when it was possible to obtain them. Nonetheless, when no significant rules were obtained there were signals of differences in the quantity of generated sequences before and after our so called "inflection point". The summarization aspect of the methodology could prove useful to experts able to interpret and expand the search horizon. Therefore, there is still room for improvement and timeframes which might be exploitable to give the user a systematic methodology to extract patterns from insider activity.

5.2 Future Work

Further work can address several improvements to this approach. One of them is the inclusion of additional quantitative information such as the price per share, overall volume information, and financial statements among others. Another feature which could enhance the descriptiveness of the patterns is the retrieving of relevant information from press releases through the application of text preprocessing techniques. This particular feature would need to be carefully designed along with subject matter experts, and it could be an independent retrieval system by itself. However, incorporating this type of information could yield more descriptive and useful rules.

Despite the work focused on enhancing the quantity of information mined patterns could contain, the fact lower support thresholds could mean too many rules for the user gives place to future work. A strategy to avoid this problem could be by reducing the temporal redundancy between the rules. In [32] the problem of deriving classification rules for sequential data are investigated. The employed classifier operates by successively refining a given pattern to better distinguish between the classes. A starting pattern is derived by those parts that are shared among all instances of the same class, then similar rules applied to some common context can should output only the rule which contains the maximum information. Then, the pattern post-processing would be enhanced to keep only the most representative sequences.

Appendix A

Dow Jones Components List

CIK	Symbol	Company Name
66740	MMM	3M CO
4281	AA	ALCOA INC
4962	AXP	AMERICAN EXPRESS CO
732717	T	AT&T INC
70858	BAC	BANK OF AMERICA CORP
12927	BA	BOEING CO
18230	CAT	CATERPILLAR INC
93410	CVX	CHEVRON CORP
858877	CSCO	CISCO SYSTEMS INC
21344	KO	COCA COLA CO
30554	DD	DUPONT E I DE NEMOURS & CO
34088	XOM	EXXON MOBIL CORP
40545	GE	GENERAL ELECTRIC CO
47217	HPQ	HEWLETT PACKARD CO
354950	HD	HOME DEPOT INC
50863	INTC	INTEL CORP
51143	IBM	INTERNATIONAL BUSINESS MACHINES CORP
200406	JNJ	JOHNSON & JOHNSON
19617	JPM	JPMORGAN CHASE & CO
1103982	KFT	KRAFT FOODS INC
63908	MCD	MCDONALDS CORP
64978	MRK	MERCK & CO INC
789019	MSFT	MICROSOFT CORP
78003	PFE	PFIZER INC
80424	PG	PROCTER & GAMBLE CO
86312	TRV	TRAVELERS COMPANIES INC.
101829	UTX	UNITED TECHNOLOGIES CORP DE
732712	VZ	VERIZON COMMUNICATIONS INC
104169	WMT	WAL-MART COM INC
1001039	DIS	WALT DISNEY CO
831001	C	CITIGROUP INC
1467858	GM	GENERAL MOTORS CORP
5272	AIG	AMERICAN INTERNATIONAL GROUP INC
764180	MO	ALTRIA GROUP INC
773840	HON	HONEYWELL INTERNATIONAL INC

Appendix B

SIC Industry List

SIC	Industry
3721	AIRCRAFT
6021	NATIONAL COMMERCIAL BANKS
2820	PLASTIC MATERIAL, SYNTH RESIN/RUBBER, CELLULOS (NO GLASS)
6199	FINANCE SERVICES
3570	COMPUTER & OFFICE EQUIPMENT
6331	FIRE, MARINE & CASUALTY INSURANCE
4813	TELEPHONE COMMUNICATIONS (NO RADIOTELEPHONE)
2670	CONVERTED PAPER & PAPERBOARD PRODS (NO CONTANER-S/BOXES)
3571	ELECTRONIC COMPUTERS
3724	AIRCRAFT ENGINES & ENGINE PARTS
2834	PHARMACEUTICAL PREPARATIONS
3350	ROLLING DRAWING & EXTRUDING OF NONFERROUS METALS
2840	SOAP, DETERGENTS, CLEANG PREPARATIONS, PERFUMES, COS-METICS
3531	CONSTRUCTION MACHINERY and EQUIP
3674	SEMICONDUCTORS & RELATED DEVICES
5812	RETAIL-EATING PLACES
8744	SERVICES-FACILITIES SUPPORT MANAGEMENT SERVICES
2911	PETROLEUM REFINING
2080	BEVERAGES
5211	RETAIL-LUMBER & OTHER BUILDING MATERIALS DEALERS
3523	FARM MACHINERY & EQUIPMENT
7990	SERVICES-MISCELLANEOUS AMUSEMENT and RECREATION
3600	ELECTRONIC & OTHER ELECTRICAL EQUIPMENT (NO COMPUTER EQUIP)
5331	RETAIL-VARIETY STORES
7372	SERVICES-PREPACKAGED SOFTWARE
2000	FOOD AND KINDRED PRODUCTS
3714	MOTOR VEHICLE PARTS & ACCESSORIES
3576	COMPUTER COMMUNICATIONS EQUIPMENT
4512	AIR TRANSPORTATION, SCHEDULED
3822	AUTO CONTROLS FOR REGULATING RESIDENTIAL & COMML EN-VIRONMENTS
6324	HOSPITAL & MEDICAL SERVICE PLANS
4911	ELECTRIC SERVICES
5621	RETAIL-WOMEN'S CLOTHING STORES
3841	SURGICAL & MEDICAL INSTRUMENTS & APPARATUS
2100	TOBACCO PRODUCTS
2111	CIGARETTES
2800	CHEMICALS & ALLIED PRODUCTS
2060	SUGAR and CONFECTIONERY PRODUCTS
3711	MOTOR VEHICLES & PASSENGER CAR BODIES

Appendix C

Item Labels

item ID	Label
I1	OFFICER; 75k-100k; -[2.5,1.0]%
I2	OFFICER; 500k-750k; -[25,10]%
I3	OFFICER; 200k-250k; -[99,25]%
I4	OFFICER; 1-10k; [0.5,1.0]%
I5	DIRECTOR/OFFICER; 300k-500k; [5,10]%
I6	OFFICER; 1M$\leq$; -[99,25]%
I7	DIRECTOR/OFFICER; 1-10k; [0.5,1.0]%
I8	OFFICER; 1-10k; -[2.5,1.0]%
I9	OFFICER; 0; [10,25]%
I10	OFFICER; 0; [25,99]%
I11	OFFICER; 0; [5,10]%
I12	DIRECTOR/OFFICER; 0; -[1.0,0.5]%
I13	DIRECTOR/OFFICER; 10k-50k; [0,0.5]%
I14	OFFICER; 1M$\leq$; -[25,10]%
I15	DIRECTOR; 200k-250k; [25,99]%
I16	DIRECTOR/OFFICER; 75k-100k; -[0.5,0]%
I17	DIRECTOR/OFFICER; 300k-500k; -[2.5,1.0]%
I18	DIRECTOR/OFFICER; 150k-175k; -[1.0,0.5]%
I19	DIRECTOR/OFFICER; 100k-125k; -[0.5,0]%
I20	DIRECTOR/OFFICER; 50k-75k; -[0.5,0]%
I21	DIRECTOR/OFFICER; 10k-50k; -[0.5,0]%
I22	DIRECTOR/OFFICER; 125k-150k; -[1.0,0.5]%
I23	OFFICER; 10k-50k; -[5.0,2.5]%
I24	DIRECTOR/OFFICER; 250k-300k; -[2.5,1.0]%

I25	OFFICER; 10k-50k; -[10,5]%
I26	DIRECTOR/OFFICER; 175k-200k; -[1.0,0.5]%
I27	DIRECTOR/OFFICER; 175k-200k; -[0.5,0]%
I28	DIRECTOR/OFFICER; 1M≤; -[5.0,2.5]%
I29	OFFICER; 750k-1M; -[5.0,2.5]%
I30	OFFICER; 1M≤; -[10,5]%
I31	OFFICER; 300k-500k; -[5.0,2.5]%
I32	DIRECTOR; 300k-500k; -100%
I33	OFFICER; 125k-150k; -[5.0,2.5]%
I34	OFFICER; 50k-75k; -[5.0,2.5]%
I35	OFFICER; 150k-175k; -[5.0,2.5]%
I36	OFFICER; 175k-200k; -[2.5,1.0]%
I37	OFFICER; 100k-125k; -[10,5]%
I38	OFFICER; 50k-75k; -[10,5]%
I39	DIRECTOR; 300k-500k; -[99,25]%
I40	OFFICER; 750k-1M; [10,25]%
I41	TENPERCENTOWNER; 750k-1M; [0,0.5]%
I42	DIRECTOR/OFFICER; 1-10k; [0,0.5]%
I43	OFFICER; 100k-125k; -[1.0,0.5]%
I44	OFFICER; 50k-75k; -[0.5,0]%
I45	OFFICER; 0; -[25,10]%
I46	OFFICER; 500k-750k; -[5.0,2.5]%
I47	DIRECTOR/OFFICER; 1M≤; -[2.5,1.0]%
I48	OFFICER; 0; 100% ≤
I49	OFFICER; 10k-50k; -[0.5,0]%
I50	OFFICER; 300k-500k; -[2.5,1.0]%
I51	OFFICER; 75k-100k; -[0.5,0]%
I52	DIRECTOR; 1-10k; [0.5,1.0]%
I53	OFFICER; 200k-250k; [10,25]%
I54	OFFICER; 0; -[1.0,0.5]%
I55	OFFICER; 50k-75k; -[25,10]%
I56	OFFICER; 500k-750k; [25,99]%
I57	DIRECTOR; 10k-50k; [0.5,1.0]%
I58	DIRECTOR; 10k-50k; [2.5,5.0]%
I59	DIRECTOR; 10k-50k; [1.0,2.5]%
I60	OFFICER; 75k-100k; -[10,5]%
I61	OFFICER; 500k-750k; 100% ≤

I62	DIRECTOR; 1-10k; [0,0.5]%
I63	OFFICER; 1-10k; [0,0.5]%
I64	TENPERCENTOWNER; 1M≤; -[0.5,0]%
I65	DIRECTOR/OFFICER; 0; [0,0.5]%
I66	OFFICER; 10k-50k; -[99,25]%
I67	OFFICER; 100k-125k; -[99,25]%
I68	OFFICER; 50k-75k; -[99,25]%
I69	OFFICER; 10k-50k; -[25,10]%
I70	OFFICER; 300k-500k; -[10,5]%
I71	DIRECTOR/OFFICER; 1M≤; -[10,5]%
I72	OFFICER; 200k-250k; -[5.0,2.5]%
I73	DIRECTOR/OFFICER; 750k-1M; -[5.0,2.5]%
I74	OFFICER; 0; [0,0.5]%
I75	DIRECTOR/OFFICER; 1M≤; -[99,25]%
I76	OFFICER; 125k-150k; -[10,5]%
I77	DIRECTOR/OFFICER; 1M≤; -[25,10]%
I78	OFFICER; 1M≤; -100%
I79	OFFICER; 300k-500k; -100%
I80	DIRECTOR; 0; -[99,25]%
I81	OFFICER; 175k-200k; -[25,10]%
I82	OFFICER; 75k-100k; -[25,10]%
I83	OFFICER; 125k-150k; -[99,25]%
I84	OFFICER; 1-10k; -[5.0,2.5]%
I85	OFFICER; 100k-125k; -[25,10]%
I86	OFFICER; 200k-250k; -[25,10]%
I87	OFFICER; 150k-175k; -[2.5,1.0]%
I88	OFFICER; 0; [0.5,1.0]%
I89	OFFICER; 100k-125k; [2.5,5.0]%
I90	OFFICER; 175k-200k; NA
I91	OFFICER; 1M≤; 100% ≤
I92	OFFICER; 75k-100k; -[99,25]%
I93	DIRECTOR/OFFICER; 500k-750k; -[2.5,1.0]%
I94	DIRECTOR; 1-10k; [1.0,2.5]%
I95	OFFICER; 10k-50k; -[2.5,1.0]%
I96	TENPERCENTOWNER; 0; [0,0.5]%
I97	OFFICER; 1-10k; -[10,5]%
I98	OFFICER; 1-10k; [5,10]%

I99	DIRECTOR/OFFICER; 0; 100% $\leq$
I100	OFFICER; 1-10k; -[1.0,0.5]%
I101	OFFICER; 1-10k; -[0.5,0]%
I102	OFFICER; 10k-50k; -[1.0,0.5]%
I103	OFFICER; 0; -[2.5,1.0]%
I104	OFFICER; 1-10k; [2.5,5.0]%
I105	OTHER; 175k-200k; -[2.5,1.0]%
I106	OFFICER; 750k-1M; -[99,25]%
I107	OTHER; 750k-1M; -[10,5]%
I108	OTHER; 500k-750k; -[10,5]%
I109	OTHER; 1-10k; -[0.5,0]%
I110	OTHER; 1M$\leq$; -[25,10]%
I111	DIRECTOR; 125k-150k; [25,99]%
I112	DIRECTOR/OFFICER; 10k-50k; [5,10]%
I113	DIRECTOR/OFFICER; 750k-1M; -[10,5]%
I114	DIRECTOR/OFFICER; 300k-500k; -[5.0,2.5]%
I115	OFFICER; 150k-175k; -[10,5]%
I116	OFFICER; 100k-125k; [25,99]%
I117	DIRECTOR/OFFICER; 500k-750k; -[10,5]%
I118	OFFICER; 10k-50k; [2.5,5.0]%
I119	OFFICER; 300k-500k; [10,25]%
I120	OFFICER; 500k-750k; [10,25]%
I121	DIRECTOR; 0; [5,10]%
I122	OFFICER; 300k-500k; -[25,10]%
I123	OFFICER; 125k-150k; -100%
I124	OFFICER; 500k-750k; -100%
I125	DIRECTOR/OFFICER; 1M$\leq$; 100% $\leq$
I126	DIRECTOR; 10k-50k; [0,0.5]%
I127	OFFICER; 1-10k; [1.0,2.5]%
I128	OFFICER; 175k-200k; -[10,5]%
I129	OFFICER; 500k-750k; -[99,25]%
I130	OFFICER; 750k-1M; -[10,5]%
I131	DIRECTOR; 10k-50k; [5,10]%
I132	OFFICER; 150k-175k; -[0.5,0]%
I133	DIRECTOR; 75k-100k; [0.5,1.0]%
I134	OFFICER; 750k-1M; -[25,10]%
I135	OFFICER; 1M$\leq$; -[5.0,2.5]%

I136	DIRECTOR; 1-10k; [2.5,5.0]%
I137	DIRECTOR; 100k-125k; -[25,10]%
I138	OTHER; 10k-50k; -[0.5,0]%
I139	DIRECTOR; 50k-75k; -[99,25]%
I140	OFFICER; 250k-300k; -[25,10]%
I141	OFFICER; 500k-750k; -[10,5]%
I142	DIRECTOR/OFFICER; 300k-500k; -[0.5,0]%
I143	DIRECTOR/OFFICER; 150k-175k; -[0.5,0]%
I144	DIRECTOR/OFFICER; 125k-150k; -[0.5,0]%
I145	OFFICER; 0; -[0.5,0]%
I146	OFFICER; 0; -[10,5]%
I147	OFFICER; 0; [1.0,2.5]%
I148	DIRECTOR/OFFICER; 0; -[0.5,0]%
I149	DIRECTOR/OFFICER; 0; [0.5,1.0]%
I150	OFFICER; 0; [2.5,5.0]%
I151	DIRECTOR/OFFICER; 750k-1M; -[1.0,0.5]%
I152	DIRECTOR; 10k-50k; [25,99]%
I153	DIRECTOR/OFFICER; 200k-250k; -[1.0,0.5]%
I154	DIRECTOR/OFFICER; 200k-250k; -[2.5,1.0]%
I155	DIRECTOR/OFFICER; 1-10k; NA
I156	OFFICER; 75k-100k; -[1.0,0.5]%
I157	OTHER; 750k-1M; -[25,10]%
I158	OFFICER; 500k-750k; -[2.5,1.0]%
I159	DIRECTOR; 50k-75k; -[2.5,1.0]%
I160	OFFICER; 50k-75k; -[1.0,0.5]%
I161	OFFICER; 200k-250k; -[2.5,1.0]%
I162	OTHER; 300k-500k; -[10,5]%
I163	DIRECTOR/OFFICER/OTHER; 500k-750k; -[5.0,2.5]%
I164	OFFICER; 0; -[5.0,2.5]%
I165	OFFICER; 10k-50k; [5,10]%
I166	OFFICER; 10k-50k; [0.5,1.0]%
I167	OFFICER; 10k-50k; [0,0.5]%
I168	OFFICER; 10k-50k; [1.0,2.5]%
I169	OFFICER; 500k-750k; -[1.0,0.5]%
I170	DIRECTOR/OFFICER; 300k-500k; -[1.0,0.5]%
I171	OFFICER; 50k-75k; -[2.5,1.0]%
I172	OFFICER; 300k-500k; -[1.0,0.5]%

I173	OFFICER; 250k-300k; -[5.0,2.5]%
I174	OFFICER; 250k-300k; -[2.5,1.0]%
I175	OTHER; 300k-500k; -[5.0,2.5]%
I176	OTHER; 100k-125k; -[5.0,2.5]%
I177	OTHER; 75k-100k; -[2.5,1.0]%
I178	DIRECTOR/OFFICER; 250k-300k; -[1.0,0.5]%
I179	DIRECTOR/OFFICER; 50k-75k; [0,0.5]%
I180	OFFICER; 250k-300k; -[10,5]%
I181	DIRECTOR/OFFICER; 75k-100k; -[1.0,0.5]%
I182	DIRECTOR/OFFICER; 500k-750k; -[5.0,2.5]%
I183	OFFICER; 125k-150k; -[0.5,0]%
I184	OFFICER; 750k-1M; -[1.0,0.5]%
I185	OFFICER; 150k-175k; -[25,10]%
I186	DIRECTOR; 500k-750k; -[25,10]%
I187	DIRECTOR; 1M≤; -[25,10]%
I188	DIRECTOR; 0; -[0.5,0]%
I189	OFFICER; 150k-175k; -[99,25]%
I190	OFFICER; 175k-200k; -[5.0,2.5]%
I191	DIRECTOR; 0; -[5.0,2.5]%
I192	DIRECTOR; 100k-125k; -[99,25]%
I193	DIRECTOR; 0; 100% ≤
I194	DIRECTOR/OFFICER; 0; [5,10]%
I195	DIRECTOR; 150k-175k; [25,99]%
I196	DIRECTOR; 10k-50k; 100% ≤
I197	OFFICER; 125k-150k; -[2.5,1.0]%
I198	DIRECTOR/OFFICER; 500k-750k; -[1.0,0.5]%
I199	OFFICER; 1M≤; -[2.5,1.0]%
I200	OFFICER; 0; -100%
I201	OFFICER; 300k-500k; -[99,25]%
I202	DIRECTOR; 0; -100%
I203	DIRECTOR; 50k-75k; 100% ≤
I204	OFFICER; 75k-100k; -[5.0,2.5]%
I205	DIRECTOR/OFFICER; 0; -[5.0,2.5]%
I206	DIRECTOR/OFFICER; 0; [10,25]%
I207	DIRECTOR/OFFICER; 300k-500k; [1.0,2.5]%
I208	OFFICER; 100k-125k; -[2.5,1.0]%
I209	OFFICER; 125k-150k; [1.0,2.5]%

I210	DIRECTOR/OFFICER; 0; [25,99]%
I211	DIRECTOR; 50k-75k; [10,25]%
I212	DIRECTOR/OFFICER; 1-10k; -[0.5,0]%
I213	OFFICER; 125k-150k; -[25,10]%
I214	DIRECTOR/OFFICER; 0; [2.5,5.0]%
I215	OFFICER; 10k-50k; -100%
I216	DIRECTOR; 200k-250k; 100% ≤
I217	OFFICER; 250k-300k; [25,99]%
I218	DIRECTOR/OFFICER; 175k-200k; -[2.5,1.0]%
I219	DIRECTOR; 50k-75k; [5,10]%
I220	DIRECTOR/OFFICER; 200k-250k; -[0.5,0]%
I221	DIRECTOR; 75k-100k; [10,25]%
I222	DIRECTOR/OFFICER; 1M≤; [10,25]%
I223	DIRECTOR; 0; [0,0.5]%
I224	DIRECTOR/OFFICER; 750k-1M; -[2.5,1.0]%
I225	TENPERCENTOWNER; 50k-75k; [0,0.5]%
I226	DIRECTOR; 10k-50k; [10,25]%
I227	OFFICER; 10k-50k; [25,99]%
I228	OFFICER; 50k-75k; [25,99]%
I229	OFFICER; 10k-50k; [10,25]%
I230	OFFICER; 100k-125k; -[5.0,2.5]%
I231	DIRECTOR/OFFICER; 1M≤; -[0.5,0]%
I232	OFFICER; 0; -[99,25]%
I233	DIRECTOR/OFFICER; 125k-150k; [1.0,2.5]%
I234	OFFICER; 150k-175k; -[1.0,0.5]%
I235	DIRECTOR; 75k-100k; -[25,10]%
I236	OFFICER; 75k-100k; [10,25]%
I237	OFFICER; 750k-1M; 100% ≤
I238	OFFICER; 250k-300k; -[99,25]%
I239	DIRECTOR; 1-10k; [25,99]%
I240	TENPERCENTOWNER; 300k-500k; -[0.5,0]%
I241	DIRECTOR/OFFICER; 175k-200k; [0,0.5]%
I242	DIRECTOR; 1-10k; 100% ≤
I243	DIRECTOR; 50k-75k; [1.0,2.5]%
I244	OFFICER; 300k-500k; 100% ≤
I245	OFFICER; 175k-200k; -[99,25]%
I246	DIRECTOR/OFFICER/TENPERCENTOWNER; 1M≤; -[0.5,0]%

I247	OFFICER; 75k-100k; [1.0,2.5]%
I248	DIRECTOR; 50k-75k; [2.5,5.0]%
I249	OFFICER; 175k-200k; -[0.5,0]%
I250	OFFICER; 300k-500k; [25,99]%
I251	DIRECTOR; 175k-200k; [10,25]%
I252	DIRECTOR; 1-10k; [5,10]%
I253	OFFICER; 100k-125k; -100%
I254	TENPERCENTOWNER; 10k-50k; [0,0.5]%
I255	DIRECTOR/OFFICER/TENPERCENTOWNER; 1M≤; [0,0.5]%
I256	DIRECTOR; 0; [2.5,5.0]%
I257	DIRECTOR; 300k-500k; 100% ≤
I258	DIRECTOR; 0; [1.0,2.5]%
I259	DIRECTOR/OFFICER; 0; -[2.5,1.0]%
I260	DIRECTOR; 100k-125k; [1.0,2.5]%
I261	DIRECTOR/OFFICER; 100k-125k; -[1.0,0.5]%
I262	DIRECTOR; 100k-125k; [25,99]%
I263	DIRECTOR; 75k-100k; [25,99]%
I264	DIRECTOR; 100k-125k; [10,25]%
I265	DIRECTOR; 100k-125k; [2.5,5.0]%
I266	OFFICER; 200k-250k; -[10,5]%
I267	DIRECTOR/OFFICER/TENPERCENTOWNER; 750k-1M; [0,0.5]%
I268	DIRECTOR/OFFICER/TENPERCENTOWNER; 500k-750k; [0,0.5]%
I269	OFFICER; 1-10k; 100% ≤
I270	DIRECTOR; 250k-300k; -[0.5,0]%
I271	DIRECTOR; 500k-750k; -[2.5,1.0]%
I272	DIRECTOR/OFFICER; 500k-750k; -[0.5,0]%
I273	OFFICER; 75k-100k; [5,10]%
I274	DIRECTOR; 100k-125k; -[10,5]%
I275	DIRECTOR; 0; [0.5,1.0]%
I276	DIRECTOR; 0; -[10,5]%
I277	DIRECTOR; 100k-125k; [5,10]%
I278	DIRECTOR; 750k-1M; -[0.5,0]%
I279	DIRECTOR; 1M≤; -[0.5,0]%
I280	OFFICER; 300k-500k; -[0.5,0]%
I281	DIRECTOR; 125k-150k; 100% ≤
I282	DIRECTOR; 10k-50k; -[10,5]%
I283	DIRECTOR; 150k-175k; -[25,10]%

I284	OFFICER; 50k-75k; [10,25]%
I285	DIRECTOR/OFFICER/TENPERCENTOWNER; 125k-150k; [0,0.5]%
I286	DIRECTOR; 0; -[25,10]%
I287	DIRECTOR; 0; [25,99]%
I288	OFFICER; 250k-300k; -[1.0,0.5]%
I289	DIRECTOR/OFFICER/TENPERCENTOWNER; 100k-125k; [0,0.5]%
I290	DIRECTOR; 750k-1M; [0.5,1.0]%
I291	DIRECTOR/OFFICER; 50k-75k; [1.0,2.5]%
I292	DIRECTOR; 1M≤; [1.0,2.5]%
I293	DIRECTOR; 1M≤; [0.5,1.0]%
I294	DIRECTOR; 1M≤; [2.5,5.0]%
I295	DIRECTOR; 50k-75k; -[25,10]%
I296	DIRECTOR; 1M≤; [5,10]%
I297	DIRECTOR; 75k-100k; [2.5,5.0]%
I298	DIRECTOR; 150k-175k; [10,25]%
I299	DIRECTOR; 125k-150k; [10,25]%
I300	OFFICER; 50k-75k; [1.0,2.5]%
I301	DIRECTOR; 125k-150k; [5,10]%
I302	DIRECTOR; 125k-150k; [2.5,5.0]%
I303	OFFICER; 10k-50k; 100% ≤
I304	DIRECTOR; 100k-125k; 100% ≤
I305	DIRECTOR/OFFICER/TENPERCENTOWNER; 1-10k; [0,0.5]%
I306	DIRECTOR/OFFICER/TENPERCENTOWNER; 300k-500k; [0,0.5]%
I307	TENPERCENTOWNER; 250k-300k; [0,0.5]%
I308	DIRECTOR; 150k-175k; [1.0,2.5]%
I309	DIRECTOR; 300k-500k; [10,25]%
I310	OFFICER; 125k-150k; -[1.0,0.5]%
I311	DIRECTOR/OFFICER/TENPERCENTOWNER; 75k-100k; [0,0.5]%
I312	DIRECTOR/OFFICER/TENPERCENTOWNER; 10k-50k; [0,0.5]%
I313	DIRECTOR/OFFICER/TENPERCENTOWNER; 250k-300k; [0,0.5]%
I314	DIRECTOR/OFFICER/TENPERCENTOWNER; 50k-75k; [0,0.5]%
I315	DIRECTOR/OFFICER/TENPERCENTOWNER; 150k-175k; [0,0.5]%
I316	DIRECTOR/OFFICER/TENPERCENTOWNER; 200k-250k; [0,0.5]%
I317	DIRECTOR/OFFICER/TENPERCENTOWNER; 175k-200k; [0,0.5]%
I318	DIRECTOR/OFFICER; 250k-300k; -[0.5,0]%
I319	DIRECTOR; 500k-750k; -[1.0,0.5]%
I320	DIRECTOR; 300k-500k; -[0.5,0]%

I321	DIRECTOR; 1M≤; -[2.5,1.0]%
I322	DIRECTOR; 750k-1M; -[2.5,1.0]%
I323	DIRECTOR/OFFICER/OTHER; 100k-125k; -[1.0,0.5]%
I324	DIRECTOR/OFFICER; 1-10k; -100%
I325	DIRECTOR; 125k-150k; -[0.5,0]%
I326	DIRECTOR; 200k-250k; -[0.5,0]%
I327	DIRECTOR; 10k-50k; -[0.5,0]%
I328	DIRECTOR; 0; [10,25]%
I329	DIRECTOR; 50k-75k; [25,99]%
I330	OFFICER/OTHER; 750k-1M; -[99,25]%
I331	OFFICER; 175k-200k; -[1.0,0.5]%
I332	DIRECTOR; 0; -[2.5,1.0]%
I333	DIRECTOR; 300k-500k; [25,99]%
I334	DIRECTOR; 0; -[1.0,0.5]%
I335	DIRECTOR; 1M≤; -[10,5]%
I336	TENPERCENTOWNER; 1-10k; [0,0.5]%
I337	DIRECTOR/OFFICER; 0; -100%
I338	OFFICER/OTHER; 125k-150k; -[10,5]%
I339	OFFICER/OTHER; 75k-100k; -[10,5]%
I340	TENPERCENTOWNER; 100k-125k; [0,0.5]%
I341	DIRECTOR/OFFICER; 50k-75k; -[1.0,0.5]%
I342	DIRECTOR; 1M≤; -[5.0,2.5]%
I343	DIRECTOR; 1-10k; [10,25]%
I344	OFFICER; 1-10k; [10,25]%
I345	DIRECTOR/OFFICER; 500k-750k; -[99,25]%
I346	OFFICER; 1-10k; [25,99]%
I347	OFFICER; 200k-250k; -100%
I348	DIRECTOR; 10k-50k; -100%
I349	OFFICER; 100k-125k; -[0.5,0]%
I350	TENPERCENTOWNER; 75k-100k; [0,0.5]%
I351	DIRECTOR/OFFICER; 175k-200k; -[5.0,2.5]%
I352	OFFICER/OTHER; 100k-125k; -[10,5]%
I353	OFFICER/OTHER; 75k-100k; -[25,10]%
I354	OFFICER/OTHER; 100k-125k; -[25,10]%
I355	DIRECTOR; 150k-175k; 100% ≤
I356	OFFICER/OTHER; 10k-50k; -[10,5]%
I357	OFFICER/OTHER; 200k-250k; -[25,10]%

I358	DIRECTOR; 200k-250k; [10,25]%
I359	OFFICER/OTHER; 1M≤; -[99,25]%
I360	DIRECTOR/OFFICER; 50k-75k; [0.5,1.0]%
I361	OFFICER/OTHER; 300k-500k; -[99,25]%
I362	OFFICER; 500k-750k; [5,10]%
I363	DIRECTOR; 50k-75k; -[10,5]%
I364	DIRECTOR/OFFICER; 0; [1.0,2.5]%
I365	DIRECTOR/OFFICER; 750k-1M; -[25,10]%
I366	OTHER; 250k-300k; -[99,25]%
I367	OTHER; 750k-1M; -[99,25]%
I368	OTHER; 500k-750k; [10,25]%
I369	OTHER; 250k-300k; [1.0,2.5]%
I370	OTHER; 500k-750k; [25,99]%
I371	OTHER; 250k-300k; [25,99]%
I372	OTHER; 1M≤; [5,10]%
I373	OTHER; 100k-125k; [0.5,1.0]%
I374	OTHER; 1M≤; [10,25]%
I375	OTHER; 1M≤; [25,99]%
I376	OTHER; 150k-175k; [1.0,2.5]%
I377	OTHER; 300k-500k; [10,25]%
I378	OTHER; 200k-250k; [1.0,2.5]%
I379	OTHER; 100k-125k; [2.5,5.0]%
I380	OTHER; 250k-300k; [10,25]%
I381	OTHER; 500k-750k; [2.5,5.0]%
I382	OTHER; 250k-300k; [5,10]%
I383	OTHER; 300k-500k; [5,10]%
I384	OTHER; 500k-750k; [5,10]%
I385	OTHER; 500k-750k; 100% ≤
I386	OTHER; 300k-500k; [2.5,5.0]%
I387	DIRECTOR/OFFICER/OTHER; 1M≤; -[1.0,0.5]%
I388	DIRECTOR/OFFICER/OTHER; 1M≤; -[0.5,0]%
I389	DIRECTOR/OFFICER/OTHER; 500k-750k; -[0.5,0]%
I390	DIRECTOR/OTHER; 1M≤; -[0.5,0]%
I391	DIRECTOR/OTHER; 750k-1M; -[0.5,0]%
I392	DIRECTOR/OTHER; 500k-750k; -[0.5,0]%
I393	DIRECTOR/OFFICER/OTHER; 250k-300k; -[0.5,0]%
I394	DIRECTOR/OFFICER/OTHER; 300k-500k; -[0.5,0]%

I395	OTHER; 200k-250k; [0.5,1.0]%
I396	DIRECTOR/OTHER; 300k-500k; -[0.5,0]%
I397	DIRECTOR/OFFICER/OTHER; 75k-100k; -[0.5,0]%
I398	DIRECTOR/OTHER; 200k-250k; -[0.5,0]%
I399	OTHER; 300k-500k; [1.0,2.5]%
I400	OTHER; 250k-300k; [0.5,1.0]%
I401	DIRECTOR/OTHER; 75k-100k; -[0.5,0]%
I402	DIRECTOR/OFFICER/OTHER; 750k-1M; -[0.5,0]%
I403	DIRECTOR/OFFICER/OTHER; 200k-250k; -[0.5,0]%
I404	DIRECTOR/OTHER; 1M≤; -[1.0,0.5]%
I405	DIRECTOR/OTHER; 250k-300k; -[0.5,0]%
I406	OTHER; 750k-1M; [2.5,5.0]%
I407	DIRECTOR; 75k-100k; [1.0,2.5]%
I408	DIRECTOR/OFFICER/OTHER; 175k-200k; -[0.5,0]%
I409	OTHER; 750k-1M; [1.0,2.5]%
I410	DIRECTOR/OFFICER/OTHER; 125k-150k; -[0.5,0]%
I411	DIRECTOR/OFFICER/OTHER; 100k-125k; -[0.5,0]%
I412	OTHER; 1M≤; [2.5,5.0]%
I413	DIRECTOR/OFFICER/OTHER; 10k-50k; [0,0.5]%
I414	DIRECTOR/OTHER; 125k-150k; -[0.5,0]%
I415	DIRECTOR/OTHER; 175k-200k; -[0.5,0]%
I416	OTHER; 500k-750k; [1.0,2.5]%
I417	DIRECTOR/OFFICER/OTHER; 10k-50k; -[0.5,0]%
I418	DIRECTOR/OTHER; 100k-125k; -[0.5,0]%
I419	DIRECTOR/OTHER; 10k-50k; -[0.5,0]%
I420	DIRECTOR/OTHER; 50k-75k; -[0.5,0]%
I421	DIRECTOR/OFFICER/OTHER; 50k-75k; -[0.5,0]%
I422	DIRECTOR/OTHER; 10k-50k; [0,0.5]%
I423	DIRECTOR/OFFICER/OTHER; 1M≤; -[10,5]%
I424	DIRECTOR/OFFICER/OTHER; 1M≤; -[2.5,1.0]%
I425	DIRECTOR/OTHER; 1M≤; -[2.5,1.0]%
I426	DIRECTOR/OTHER; 1M≤; -[10,5]%
I427	DIRECTOR/OFFICER/OTHER; 1M≤; -[25,10]%
I428	DIRECTOR/OFFICER/OTHER; 750k-1M; -[1.0,0.5]%
I429	DIRECTOR/OTHER; 500k-750k; -[1.0,0.5]%
I430	DIRECTOR/OTHER; 750k-1M; -[1.0,0.5]%
I431	DIRECTOR/OFFICER/OTHER; 500k-750k; -[1.0,0.5]%

I432	DIRECTOR/OTHER; 1-10k; [0,0.5]%
I433	DIRECTOR/OFFICER/OTHER; 1-10k; [0,0.5]%
I434	DIRECTOR/OTHER; 1M≤; -[25,10]%
I435	DIRECTOR/OFFICER/OTHER; 1M≤; -[5.0,2.5]%
I436	DIRECTOR/OTHER; 1M≤; -[5.0,2.5]%
I437	DIRECTOR/OTHER; 300k-500k; -[1.0,0.5]%
I438	DIRECTOR/OFFICER/OTHER; 750k-1M; -[2.5,1.0]%
I439	DIRECTOR/OFFICER/OTHER; 300k-500k; -[1.0,0.5]%
I440	DIRECTOR/OTHER; 750k-1M; -[2.5,1.0]%
I441	DIRECTOR/OFFICER/OTHER; 150k-175k; -[0.5,0]%
I442	DIRECTOR/OTHER; 150k-175k; -[0.5,0]%
I443	DIRECTOR/OTHER; 1M≤; -[99,25]%
I444	DIRECTOR/OTHER; 250k-300k; -[99,25]%
I445	DIRECTOR/OTHER; 300k-500k; -[99,25]%
I446	DIRECTOR/OTHER; 300k-500k; -[25,10]%
I447	DIRECTOR/OTHER; 250k-300k; -100%
I448	DIRECTOR/OTHER; 500k-750k; -[99,25]%
I449	DIRECTOR/OFFICER/OTHER; 500k-750k; -[10,5]%
I450	DIRECTOR/OFFICER/OTHER; 500k-750k; -[2.5,1.0]%
I451	OFFICER; 100k-125k; [10,25]%
I452	OFFICER; 150k-175k; -100%
I453	OFFICER; 150k-175k; [1.0,2.5]%
I454	OFFICER; 125k-150k; 100% ≤
I455	DIRECTOR; 75k-100k; 100% ≤
I456	DIRECTOR/OFFICER/OTHER; 300k-500k; -[10,5]%
I457	DIRECTOR/OFFICER/OTHER; 200k-250k; -[1.0,0.5]%
I458	OTHER; 50k-75k; -[5.0,2.5]%
I459	OFFICER; 50k-75k; -100%
I460	DIRECTOR; 1-10k; -[0.5,0]%
I461	OFFICER; 75k-100k; -100%
I462	OFFICER; 1-10k; -100%
I463	DIRECTOR/OFFICER; 1M≤; -[1.0,0.5]%
I464	DIRECTOR; 500k-750k; [10,25]%
I465	DIRECTOR/OFFICER/OTHER; 300k-500k; -[2.5,1.0]%
I466	DIRECTOR/OFFICER/OTHER; 300k-500k; -[25,10]%
I467	DIRECTOR/OFFICER/OTHER; 150k-175k; -[1.0,0.5]%
I468	DIRECTOR/OFFICER/OTHER; 1M≤; -[99,25]%

I469	DIRECTOR/OFFICER; 100k-125k; -[2.5,1.0]%
I470	DIRECTOR/OFFICER/OTHER; 175k-200k; -[1.0,0.5]%
I471	DIRECTOR/OFFICER/OTHER; 500k-750k; -[99,25]%
I472	OFFICER; 175k-200k; [10,25]%
I473	DIRECTOR/OFFICER; 0; -[25,10]%
I474	DIRECTOR; 1M≤; 100% ≤
I475	DIRECTOR; 1-10k; -[2.5,1.0]%
I476	DIRECTOR; 10k-50k; -[5.0,2.5]%
I477	DIRECTOR; 500k-750k; -[99,25]%
I478	DIRECTOR/OFFICER/OTHER; 300k-500k; -[99,25]%
I479	DIRECTOR; 1M≤; [25,99]%
I480	OFFICER; 75k-100k; [25,99]%
I481	OTHER; 0; -[1.0,0.5]%
I482	DIRECTOR/OTHER; 500k-750k; -[2.5,1.0]%
I483	DIRECTOR/OTHER; 200k-250k; -[1.0,0.5]%
I484	DIRECTOR/OFFICER/OTHER; 250k-300k; -100%
I485	DIRECTOR/OFFICER/OTHER; 250k-300k; -[99,25]%
I486	OFFICER; 150k-175k; [25,99]%
I487	DIRECTOR/OTHER; 300k-500k; -[2.5,1.0]%
I488	DIRECTOR/OTHER; 150k-175k; -[1.0,0.5]%
I489	DIRECTOR/OTHER; 300k-500k; -[5.0,2.5]%
I490	DIRECTOR/OTHER; 500k-750k; -[10,5]%
I491	DIRECTOR/OTHER; 750k-1M; -[10,5]%
I492	DIRECTOR/OTHER; 175k-200k; -[1.0,0.5]%
I493	DIRECTOR/OTHER; 100k-125k; -[1.0,0.5]%
I494	DIRECTOR/OTHER; 500k-750k; -[5.0,2.5]%
I495	OTHER; 0; -[99,25]%
I496	DIRECTOR/OTHER; 200k-250k; -[5.0,2.5]%
I497	DIRECTOR/OTHER; 300k-500k; -[10,5]%
I498	DIRECTOR/OFFICER/OTHER; 200k-250k; -[5.0,2.5]%
I499	DIRECTOR/OFFICER/OTHER; 750k-1M; -[10,5]%
I500	DIRECTOR/OFFICER/OTHER; 300k-500k; -[5.0,2.5]%
I501	DIRECTOR; 175k-200k; -[0.5,0]%
I502	DIRECTOR; 50k-75k; -[0.5,0]%
I503	DIRECTOR; 150k-175k; -[0.5,0]%
I504	DIRECTOR; 300k-500k; -[1.0,0.5]%
I505	DIRECTOR; 750k-1M; -[25,10]%

I506	OFFICER; 750k-1M; [25,99]%
I507	OFFICER; 150k-175k; 100% $\leq$
I508	OFFICER; 750k-1M; [5,10]%
I509	DIRECTOR/OFFICER; 150k-175k; -[2.5,1.0]%
I510	DIRECTOR/OFFICER; 300k-500k; -[10,5]%
I511	DIRECTOR; 1-10k; -[1.0,0.5]%
I512	OFFICER/OTHER; 500k-750k; -[25,10]%
I513	DIRECTOR; 1M$\leq$; -[99,25]%
I514	DIRECTOR/OFFICER; 500k-750k; [10,25]%
I515	DIRECTOR/OFFICER; 75k-100k; -[2.5,1.0]%
I516	OFFICER/OTHER; 1-10k; -[5.0,2.5]%
I517	OFFICER/OTHER; 1-10k; -[1.0,0.5]%
I518	DIRECTOR; 1-10k; -[5.0,2.5]%
I519	OTHER; 0; -100%
I520	OTHER; 150k-175k; -100%
I521	DIRECTOR; 1-10k; -[10,5]%
I522	DIRECTOR; 125k-150k; -[5.0,2.5]%
I523	DIRECTOR/OFFICER; 200k-250k; [0.5,1.0]%
I524	DIRECTOR/OFFICER; 100k-125k; [0,0.5]%
I525	DIRECTOR; 300k-500k; -[25,10]%
I526	OFFICER; 100k-125k; 100% $\leq$
I527	DIRECTOR/OFFICER; 125k-150k; [0.5,1.0]%
I528	OFFICER; 175k-200k; -100%
I529	DIRECTOR; 300k-500k; [5,10]%
I530	OFFICER/OTHER; 1-10k; [0,0.5]%
I531	OFFICER; 125k-150k; [25,99]%
I532	OFFICER/OTHER; 125k-150k; -[25,10]%
I533	DIRECTOR; 10k-50k; -[25,10]%
I534	DIRECTOR; 1M$\leq$; -100%
I535	DIRECTOR; 10k-50k; -[99,25]%
I536	DIRECTOR/OFFICER; 100k-125k; [1.0,2.5]%
I537	OFFICER; 750k-1M; -[2.5,1.0]%
I538	OFFICER; 50k-75k; [2.5,5.0]%
I539	DIRECTOR; 100k-125k; [0,0.5]%
I540	DIRECTOR; 100k-125k; -100%
I541	OFFICER; 50k-75k; [5,10]%
I542	DIRECTOR/OFFICER; 150k-175k; [1.0,2.5]%

I543	DIRECTOR; 250k-300k; -[99,25]%
I544	DIRECTOR/OFFICER; 750k-1M; [25,99]%
I545	OFFICER/OTHER; 200k-250k; -[10,5]%
I546	OFFICER; 75k-100k; [2.5,5.0]%
I547	DIRECTOR; 175k-200k; [5,10]%
I548	DIRECTOR; 750k-1M; -[5.0,2.5]%
I549	OFFICER; 200k-250k; -[1.0,0.5]%
I550	OFFICER; 500k-750k; [1.0,2.5]%
I551	DIRECTOR/OFFICER; 500k-750k; [0,0.5]%
I552	DIRECTOR/OFFICER; 1M≤; [0,0.5]%
I553	DIRECTOR/OFFICER; 200k-250k; [0,0.5]%
I554	DIRECTOR; 100k-125k; -[5.0,2.5]%
I555	DIRECTOR; 175k-200k; [25,99]%
I556	DIRECTOR/OFFICER; 750k-1M; [0,0.5]%
I557	DIRECTOR/OFFICER; 125k-150k; [0,0.5]%
I558	DIRECTOR; 200k-250k; -[2.5,1.0]%
I559	DIRECTOR; 500k-750k; -[5.0,2.5]%
I560	DIRECTOR; 100k-125k; -[1.0,0.5]%
I561	DIRECTOR; 750k-1M; 100% ≤
I562	DIRECTOR/OFFICER; 300k-500k; [0,0.5]%
I563	DIRECTOR/OFFICER; 150k-175k; [0,0.5]%
I564	DIRECTOR; 1M≤; [10,25]%
I565	DIRECTOR/OFFICER; 10k-50k; [25,99]%
I566	OFFICER; 1M≤; [10,25]%
I567	DIRECTOR; 100k-125k; -[0.5,0]%
I568	OFFICER; 250k-300k; [10,25]%
I569	OFFICER; 300k-500k; [5,10]%
I570	DIRECTOR; 50k-75k; -100%
I571	DIRECTOR; 175k-200k; -[5.0,2.5]%
I572	DIRECTOR; 250k-300k; -[2.5,1.0]%
I573	DIRECTOR/OFFICER; 250k-300k; [0,0.5]%
I574	DIRECTOR; 175k-200k; -[10,5]%
I575	DIRECTOR; 250k-300k; [5,10]%
I576	OTHER; 500k-750k; -[99,25]%
I577	OTHER; 0; [5,10]%
I578	DIRECTOR; 500k-750k; [25,99]%
I579	DIRECTOR/OFFICER; 50k-75k; -[99,25]%

I580	DIRECTOR; 10k-50k; -[2.5,1.0]%
I581	TENPERCENTOWNER; 150k-175k; [0,0.5]%
I582	DIRECTOR; 500k-750k; [0,0.5]%
I583	DIRECTOR; 75k-100k; -[0.5,0]%
I584	DIRECTOR; 175k-200k; -[1.0,0.5]%
I585	DIRECTOR; 300k-500k; -[2.5,1.0]%
I586	DIRECTOR; 125k-150k; -[1.0,0.5]%
I587	OTHER; 50k-75k; [5,10]%
I588	OTHER; 100k-125k; -[10,5]%
I589	DIRECTOR/OFFICER; 500k-750k; -[25,10]%
I590	DIRECTOR; 200k-250k; [2.5,5.0]%
I591	DIRECTOR; 50k-75k; [0.5,1.0]%
I592	OFFICER; 750k-1M; -100%
I593	DIRECTOR/OFFICER; 10k-50k; -[1.0,0.5]%
I594	DIRECTOR; 125k-150k; -100%
I595	DIRECTOR; 125k-150k; -[25,10]%
I596	DIRECTOR; 125k-150k; [0,0.5]%
I597	DIRECTOR; 75k-100k; -[10,5]%
I598	DIRECTOR/OFFICER; 250k-300k; -[5.0,2.5]%
I599	TENPERCENTOWNER; 0; 100% $\leq$
I600	TENPERCENTOWNER; 0; -[2.5,1.0]%
I601	DIRECTOR/OFFICER; 10k-50k; [0.5,1.0]%
I602	DIRECTOR; 125k-150k; -[10,5]%
I603	TENPERCENTOWNER; 0; [2.5,5.0]%
I604	TENPERCENTOWNER; 0; [1.0,2.5]%
I605	TENPERCENTOWNER; 0; [25,99]%
I606	DIRECTOR/OFFICER; 50k-75k; -[2.5,1.0]%
I607	DIRECTOR/OFFICER; 150k-175k; -[5.0,2.5]%
I608	OFFICER; 200k-250k; [25,99]%
I609	OFFICER; 250k-300k; [5,10]%
I610	OFFICER; 200k-250k; [2.5,5.0]%
I611	OTHER; 1M$\leq$; -[10,5]%
I612	OTHER; 1M$\leq$; -[5.0,2.5]%
I613	TENPERCENTOWNER; 0; -[99,25]%
I614	TENPERCENTOWNER; 0; -[25,10]%
I615	OTHER; 1M$\leq$; -[2.5,1.0]%
I616	DIRECTOR/OFFICER; 125k-150k; -[5.0,2.5]%

I617	DIRECTOR; 200k-250k; -[99,25]%
I618	DIRECTOR/OFFICER; 0; -[10,5]%
I619	DIRECTOR/OFFICER; 100k-125k; -[5.0,2.5]%
I620	DIRECTOR/OFFICER; 10k-50k; -[2.5,1.0]%
I621	DIRECTOR/OFFICER; 200k-250k; -[5.0,2.5]%
I622	DIRECTOR/OFFICER; 75k-100k; [1.0,2.5]%
I623	DIRECTOR/OFFICER; 300k-500k; -[25,10]%
I624	OFFICER; 175k-200k; [1.0,2.5]%
I625	DIRECTOR/OFFICER; 200k-250k; [1.0,2.5]%
I626	DIRECTOR/OFFICER; 10k-50k; 100% $\leq$
I627	DIRECTOR; 75k-100k; -[5.0,2.5]%
I628	DIRECTOR; 200k-250k; -[25,10]%
I629	DIRECTOR/OFFICER; 75k-100k; [0,0.5]%
I630	DIRECTOR/OFFICER; 125k-150k; -[2.5,1.0]%
I631	DIRECTOR; 100k-125k; -[2.5,1.0]%
I632	DIRECTOR; 150k-175k; -[99,25]%
I633	DIRECTOR; 75k-100k; -[1.0,0.5]%
I634	DIRECTOR; 750k-1M; -[10,5]%
I635	DIRECTOR; 300k-500k; -[5.0,2.5]%
I636	DIRECTOR/OFFICER; 1M$\leq$; [25,99]%
I637	DIRECTOR/OFFICER; 0; -[99,25]%
I638	DIRECTOR; 10k-50k; -[1.0,0.5]%
I639	DIRECTOR; 750k-1M; [25,99]%
I640	OTHER; 10k-50k; [25,99]%
I641	DIRECTOR; 750k-1M; [10,25]%
I642	DIRECTOR/OFFICER; 500k-750k; [1.0,2.5]%
I643	OFFICER; 100k-125k; NA
I644	DIRECTOR/OFFICER; 75k-100k; -[99,25]%
I645	DIRECTOR/OFFICER; 1-10k; [5,10]%
I646	DIRECTOR; 750k-1M; [0,0.5]%
I647	DIRECTOR; 125k-150k; [0.5,1.0]%
I648	DIRECTOR/OFFICER; 150k-175k; -100%
I649	DIRECTOR/OFFICER; 1M$\leq$; -100%
I650	OFFICER; 1-10k; -[99,25]%
I651	OTHER; 250k-300k; 100% $\leq$
I652	DIRECTOR/OFFICER; 200k-250k; -[10,5]%
I653	DIRECTOR; 75k-100k; -100%

I654	DIRECTOR; 75k-100k; -[99,25]%
I655	DIRECTOR; 1-10k; -[25,10]%
I656	DIRECTOR; 200k-250k; -[1.0,0.5]%
I657	OTHER; 10k-50k; [0,0.5]%
I658	DIRECTOR/OFFICER; 750k-1M; -[99,25]%
I659	TENPERCENTOWNER; 175k-200k; [0,0.5]%
I660	DIRECTOR; 500k-750k; -100%
I661	OFFICER; 500k-750k; [2.5,5.0]%
I662	DIRECTOR/OFFICER; 750k-1M; [2.5,5.0]%
I663	DIRECTOR; 1-10k; -100%
I664	DIRECTOR/OFFICER; 10k-50k; [10,25]%
I665	DIRECTOR; 250k-300k; [0,0.5]%
I666	DIRECTOR; 1M≤; [0,0.5]%
I667	DIRECTOR; 125k-150k; -[2.5,1.0]%
I668	DIRECTOR; 75k-100k; [0,0.5]%
I669	DIRECTOR; 175k-200k; [0,0.5]%
I670	DIRECTOR; 300k-500k; [0,0.5]%
I671	DIRECTOR; 150k-175k; [0,0.5]%
I672	DIRECTOR; 200k-250k; [0,0.5]%
I673	TENPERCENTOWNER; 1M≤; [0,0.5]%
I674	TENPERCENTOWNER; 10k-50k; -[0.5,0]%
I675	TENPERCENTOWNER; 175k-200k; -[0.5,0]%
I676	TENPERCENTOWNER; 500k-750k; -[0.5,0]%
I677	TENPERCENTOWNER; 750k-1M; -[0.5,0]%
I678	TENPERCENTOWNER; 50k-75k; -[0.5,0]%
I679	TENPERCENTOWNER; 1M≤; -[1.0,0.5]%
I680	TENPERCENTOWNER; 0; -100%
I681	DIRECTOR/OFFICER; 100k-125k; -[99,25]%
I682	TENPERCENTOWNER; 125k-150k; -[0.5,0]%
I683	TENPERCENTOWNER; 75k-100k; -[0.5,0]%
I684	TENPERCENTOWNER; 100k-125k; -[0.5,0]%
I685	TENPERCENTOWNER; 250k-300k; -[0.5,0]%
I686	TENPERCENTOWNER; 0; -[0.5,0]%
I687	DIRECTOR; 175k-200k; -[2.5,1.0]%
I688	DIRECTOR; 250k-300k; -[25,10]%
I689	DIRECTOR; 750k-1M; -[99,25]%
I690	DIRECTOR; 125k-150k; -[99,25]%

I691	DIRECTOR; 250k-300k; -[5.0,2.5]%
I692	DIRECTOR; 300k-500k; [2.5,5.0]%
I693	DIRECTOR/OFFICER; 1M≤; [5,10]%
I694	DIRECTOR; 50k-75k; -[5.0,2.5]%
I695	DIRECTOR; 125k-150k; [1.0,2.5]%
I696	DIRECTOR; 200k-250k; [1.0,2.5]%
I697	DIRECTOR; 50k-75k; [0,0.5]%
I698	DIRECTOR; 50k-75k; -[1.0,0.5]%
I699	DIRECTOR; 75k-100k; -[2.5,1.0]%
I700	DIRECTOR; 500k-750k; -[10,5]%
I701	DIRECTOR; 200k-250k; -[5.0,2.5]%
I702	DIRECTOR/OTHER; 10k-50k; [25,99]%
I703	DIRECTOR; 150k-175k; -[2.5,1.0]%
I704	OFFICER; 250k-300k; -[0.5,0]%
I705	DIRECTOR; 250k-300k; -[1.0,0.5]%
I706	DIRECTOR/OFFICER; 250k-300k; [1.0,2.5]%
I707	DIRECTOR/TENPERCENTOWNER; 0; [0,0.5]%
I708	OFFICER; 125k-150k; [10,25]%
I709	OTHER; 200k-250k; -[5.0,2.5]%
I710	OTHER; 300k-500k; -[25,10]%
I711	OTHER; 1M≤; -[99,25]%
I712	OTHER; 10k-50k; -[1.0,0.5]%
I713	OFFICER; 250k-300k; -100%
I714	DIRECTOR/OFFICER; 300k-500k; [2.5,5.0]%
I715	DIRECTOR; 150k-175k; -[1.0,0.5]%
I716	OTHER; 200k-250k; -[99,25]%
I717	OFFICER; 50k-75k; 100% ≤
I718	DIRECTOR/OFFICER; 150k-175k; -[10,5]%
I719	DIRECTOR/OFFICER; 10k-50k; -[5.0,2.5]%
I720	DIRECTOR/OFFICER; 250k-300k; -[25,10]%
I721	DIRECTOR/OFFICER; 1M≤; [2.5,5.0]%
I722	OTHER; 150k-175k; -[5.0,2.5]%
I723	OFFICER; 175k-200k; [25,99]%
I724	OFFICER; 1M≤; [25,99]%
I725	DIRECTOR/OFFICER; 100k-125k; [0.5,1.0]%
I726	OTHER; 200k-250k; [2.5,5.0]%
I727	TENPERCENTOWNER; 1M≤; -[2.5,1.0]%

I728	OFFICER; 1M≤; -[1.0,0.5]%
I729	DIRECTOR; 250k-300k; -[10,5]%
I730	DIRECTOR; 500k-750k; [5,10]%
I731	OFFICER; 300k-500k; [2.5,5.0]%
I732	OFFICER; 150k-175k; [5,10]%
I733	OFFICER; 125k-150k; [2.5,5.0]%
I734	DIRECTOR/OFFICER; 0; NA
I735	DIRECTOR; 500k-750k; 100% ≤
I736	TENPERCENTOWNER; 1M≤; -[99,25]%
I737	TENPERCENTOWNER; 1M≤; -[5.0,2.5]%
I738	DIRECTOR/OFFICER; 300k-500k; [0.5,1.0]%
I739	DIRECTOR/OFFICER; 10k-50k; -[25,10]%
I740	DIRECTOR; 250k-300k; 100% ≤
I741	OFFICER; 100k-125k; [5,10]%
I742	OFFICER; 175k-200k; 100% ≤
I743	DIRECTOR; 150k-175k; -[10,5]%
I744	DIRECTOR/OFFICER; 300k-500k; -[99,25]%
I745	OFFICER; 75k-100k; 100% ≤
I746	DIRECTOR/OFFICER; 500k-750k; [25,99]%
I747	DIRECTOR/OFFICER; 200k-250k; [10,25]%
I748	OFFICER; 50k-75k; [0.5,1.0]%
I749	DIRECTOR/OFFICER; 75k-100k; [0.5,1.0]%
I750	DIRECTOR; 200k-250k; -[10,5]%
I751	OTHER; 0; 100% ≤
I752	OTHER; 0; -[2.5,1.0]%
I753	OTHER; 10k-50k; [5,10]%
I754	DIRECTOR; 150k-175k; [5,10]%
I755	DIRECTOR; 150k-175k; [2.5,5.0]%
I756	DIRECTOR/OFFICER; 50k-75k; -[5.0,2.5]%
I757	DIRECTOR/OFFICER; 175k-200k; -[10,5]%
I758	DIRECTOR/OFFICER; 1-10k; -[1.0,0.5]%
I759	DIRECTOR/OFFICER; 125k-150k; -[10,5]%
I760	DIRECTOR/OFFICER; 175k-200k; -[99,25]%
I761	DIRECTOR/OFFICER; 175k-200k; -[25,10]%
I762	DIRECTOR/OFFICER; 250k-300k; -[10,5]%
I763	DIRECTOR/OFFICER; 200k-250k; -[25,10]%
I764	DIRECTOR/OFFICER; 10k-50k; -100%

I765	DIRECTOR/OFFICER; 100k-125k; -[10,5]%
I766	DIRECTOR/OFFICER; 750k-1M; [1.0,2.5]%
I767	DIRECTOR; 250k-300k; [10,25]%
I768	DIRECTOR; 75k-100k; [5,10]%
I769	TENPERCENTOWNER; 100k-125k; -100%
I770	TENPERCENTOWNER; 300k-500k; -100%
I771	TENPERCENTOWNER; 1M$\leq$; -[25,10]%
I772	OFFICER; 1-10k; -[25,10]%
I773	DIRECTOR; 200k-250k; [5,10]%
I774	OFFICER; 250k-300k; 100% $\leq$
I775	OFFICER; 200k-250k; [5,10]%
I776	DIRECTOR/OFFICER; 300k-500k; [25,99]%
I777	OFFICER; 175k-200k; [5,10]%
I778	OFFICER; 150k-175k; [10,25]%
I779	DIRECTOR/OFFICER; 750k-1M; 100% $\leq$
I780	DIRECTOR/OFFICER; 200k-250k; [5,10]%
I781	DIRECTOR/OTHER; 100k-125k; -[10,5]%
I782	DIRECTOR/OFFICER; 200k-250k; -[99,25]%
I783	DIRECTOR/OFFICER; 125k-150k; [5,10]%
I784	DIRECTOR/OFFICER; 150k-175k; [10,25]%
I785	DIRECTOR; 250k-300k; [25,99]%
I786	DIRECTOR/OFFICER; 10k-50k; [1.0,2.5]%
I787	DIRECTOR/OFFICER; 75k-100k; [2.5,5.0]%
I788	DIRECTOR/OFFICER; 10k-50k; [2.5,5.0]%
I789	DIRECTOR/OFFICER; 500k-750k; [5,10]%
I790	DIRECTOR/OFFICER; 1-10k; -[10,5]%
I791	DIRECTOR/OFFICER; 250k-300k; [2.5,5.0]%
I792	DIRECTOR/OFFICER; 200k-250k; [2.5,5.0]%
I793	OFFICER; 150k-175k; [2.5,5.0]%
I794	OFFICER; 175k-200k; [2.5,5.0]%
I795	DIRECTOR/OFFICER/TENPERCENTOWNER; 0; 100% $\leq$
I796	DIRECTOR/OFFICER/TENPERCENTOWNER; 0; -[25,10]%
I797	DIRECTOR/OFFICER; 100k-125k; -100%
I798	DIRECTOR; 1M$\leq$; -[1.0,0.5]%
I799	DIRECTOR; 300k-500k; -[10,5]%
I800	OTHER; 0; [25,99]%
I801	DIRECTOR; 1-10k; -[99,25]%

I802	OFFICER; 200k-250k; -[0.5,0]%
I803	DIRECTOR/OFFICER; 100k-125k; [2.5,5.0]%
I804	DIRECTOR/OFFICER; 75k-100k; [10,25]%
I805	DIRECTOR/OFFICER; 500k-750k; 100% $\leq$
I806	OFFICER; 200k-250k; 100% $\leq$
I807	DIRECTOR/OFFICER; 1-10k; -[2.5,1.0]%
I808	DIRECTOR; 175k-200k; -[99,25]%
I809	OTHER; 150k-175k; -[10,5]%
I810	DIRECTOR/OFFICER; 150k-175k; [2.5,5.0]%
I811	DIRECTOR/OFFICER; 75k-100k; -[10,5]%
I812	DIRECTOR/OFFICER; 75k-100k; [5,10]%
I813	DIRECTOR; 500k-750k; -[0.5,0]%
I814	OFFICER; 75k-100k; NA
I815	DIRECTOR/TENPERCENTOWNER; 1M$\leq$; -[10,5]%
I816	DIRECTOR/TENPERCENTOWNER; 1M$\leq$; -[99,25]%
I817	OFFICER; 1M$\leq$; [5,10]%
I818	TENPERCENTOWNER/OTHER; 1M$\leq$; -[5.0,2.5]%
I819	TENPERCENTOWNER/OTHER; 1M$\leq$; -[25,10]%
I820	DIRECTOR/OFFICER; 1M$\leq$; [0.5,1.0]%
I821	DIRECTOR; 175k-200k; 100% $\leq$
I822	DIRECTOR/OFFICER; 150k-175k; [25,99]%
I823	DIRECTOR; 0; NA
I824	DIRECTOR/OFFICER; 250k-300k; -[99,25]%
I825	OFFICER; 50k-75k; [0,0.5]%
I826	OTHER; 500k-750k; -[25,10]%
I827	DIRECTOR; 175k-200k; -[25,10]%
I828	OTHER; 50k-75k; -[2.5,1.0]%
I829	OFFICER; 75k-100k; [0.5,1.0]%
I830	DIRECTOR; 175k-200k; -100%
I831	DIRECTOR/OFFICER; 175k-200k; [1.0,2.5]%
I832	OTHER; 10k-50k; -[5.0,2.5]%
I833	OTHER; 0; [10,25]%
I834	DIRECTOR/OFFICER; 250k-300k; [10,25]%
I835	DIRECTOR/OFFICER; 125k-150k; -[25,10]%
I836	DIRECTOR/OTHER; 10k-50k; -[5.0,2.5]%
I837	DIRECTOR/OFFICER; 1-10k; [25,99]%
I838	OFFICER; 0; NA

I839	DIRECTOR/OTHER; 750k-1M; 100% $\leq$
I840	DIRECTOR/OTHER; 750k-1M; [25,99]%
I841	DIRECTOR/OTHER; 750k-1M; [10,25]%
I842	DIRECTOR/OFFICER; 75k-100k; -[5.0,2.5]%

Bibliography

[1] M.J. Zaki. Spade: An efficient algorithm for mining frequent sequences. *Machine Learning*, 42(1):31–60, 2001.

[2] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

[3] Christian Buchta, Michael Hahsler, and with contributions from Daniel Diaz. *arulesSequences: Mining frequent sequences*, 2012. URL `http://CRAN.R-project.org/package=arulesSequences`. R package version 0.2-1.

[4] J. Lakonishok and I. Lee. Are insider trades informative? *Review of financial studies*, 14(1):79–11, 2001.

[5] L.M. Ausubel. Insider trading in a rational expectations economy. *The American Economic Review*, pages 1022–1041, 1990.

[6] S. Lustgarten and V. Mande. Financial analysts' earnings forecasts and insider trading. *Journal of Accounting and Public Policy*, 14(3):233–261, 1995.

[7] H. Goldberg, J.D. Kirkland, D. Lee, P. Shyr, and D. Thakker. The nasd securities observation, news analysis and regulation system (sonar). In *Proc. of IAAI03*, 2003.

[8] J.H. Lorie and V. Niederhoffer. Predictive and statistical properties of insider trading. *JL & Econ.*, 11:35, 1968.

[9] J.F. Jaffe. Special information and insider trading. *The Journal of Business*, 47(3): 410–428, 1974.

[10] J.E. Finnerty. Insiders and market efficiency. *The Journal of Finance*, 31(4):1141–1148, 1976.

[11] H.N. Seyhun. Overreaction or fundamentals: Some lessons from insiders' response to the market crash of 1987. *Journal of Finance*, pages 1363–1388, 1990.

[12] A. Kolasinski and X. Li. Are corporate managers savvy about their stock price? evidence from insider trading after earnings announcements. *Journal of Accounting and Public Policy*, 29(1):27–44, 2010.

[13] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD Record*, volume 22, pages 207–216. ACM, 1993.

[14] M. Houtsma and A. Swami. Set-oriented mining for association rules in relational databases. In *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*, pages 25–33. IEEE, 1995.

[15] R. Agrawal, R. Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, volume 1215, pages 487–499, 1994.

[16] M.J. Zaki, S. Parthasarathy, M. Ogihara, W. Li, et al. New algorithms for fast discovery of association rules. In *3rd Intl. Conf. on Knowledge Discovery and Data Mining*, volume 20, pages 283–286, 1997.

[17] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *ACM SIGMOD Record*, volume 29, pages 1–12. ACM, 2000.

[18] R.C. Agarwal, C.C. Aggarwal, and VVV Prasad. A tree projection algorithm for generation of frequent item sets. *Journal of parallel and Distributed Computing*, 61 (3):350–371, 2001.

[19] C. Borgelt. Keeping things simple: finding frequent item sets by recursive elimination. In *Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations*, pages 66–70. ACM, 2005.

[20] R. Agrawal and R. Srikant. Mining sequential patterns. In *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*, pages 3–14. IEEE, 1995.

[21] J. Ayres, J. Flannick, J. Gehrke, and T. Yiu. Sequential pattern mining using a bitmap representation. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 429–435. ACM, 2002.

[22] M. Garofalakis, R. Rastogi, and K. Shim. Spirit: Sequential pattern mining with regular expression constraints. In *Proceedings of the international conference on very large data bases*, pages 223–234, 1999.

[23] R. Srikant and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements. *Advances in Database TechnologyEDBT'96*, pages 1–17, 1996.

[24] C. Phua, V. Lee, K. Smith, and R. Gayler. A comprehensive survey of data mining-based fraud detection research. *Arxiv preprint arXiv:1009.6119*, 2010.

[25] Christian Borgelt. Frequent pattern mining. URL `http://www.borgelt.net/courses.html#fpm`.

[26] Mohammed J. Zaki and Limsoon Wong. Data mining techniques. In Limsoon Wong and Louxin Zhang, editors, *Selected Topics in Post-Genome Knowledge Discovery*, pages 125–163. World Scientific Publishers, 2004.

[27] M.J. Zaki. Sequence mining in categorical domains: incorporating constraints. In *Proceedings of the ninth international conference on Information and knowledge management*, pages 422–429. ACM, 2000.

[28] J. Ahola. Mining sequential patterns. *VTT research report TTE1-2001*, 10, 2001.

[29] Y. Tse. Price discovery and volatility spillovers in the djia index and futures markets. *Journal of Futures markets*, 19(8):911–930, 1999.

[30] Duncan Temple Lang. *XML: Tools for parsing and generating XML within R and S-Plus.*, 2012. URL `http://CRAN.R-project.org/package=XML`. R package version 3.9-4.

[31] Hadley Wickham. The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1):1–29, 2011. URL `http://www.jstatsoft.org/v40/i01/`.

[32] F. Höppner, S. Sebastian Peter, and M. Berthold. Enriching multivariate temporal patterns with context information to support classification. 2012.