Selected Results from Clustering and Analyzing Stock Market Trade Data

by

Zhihan Zhang

B.A., South China Agricultural University, 2016

A REPORT

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2018

Approved by:

Major Professor
Michael Higgins

Copyright

Abstract

The amount of data generated from stock market trading is massive. For example, roughly 10 million trades are performed each day on the NASDAQ stock exchange. A significant proportion of these trades are made by high-frequency traders. These entities make on the order of thousands or more trades a day. However, the stock-market factors that drive the decisions of high-frequency traders are poorly understood. Recently, hybridized threshold clustering (HTC) has been proposed as a way of clustering large-to-massive datasets. In this report, we use three months of NASDAQ HFT data—a dataset containing information on all trades of 120 different stocks including identifiers on whether the buyer and/or seller were high-frequency traders—to investigate the trading patterns of high-frequency traders, and we explore the use of HTC to identify these patterns. We find that, while HTC can be successfully performed on the NASDAQ HFT dataset, the amount of information gleaned from this clustering is limited. Instead, we show that an understanding of the habits of high-frequency traders may be gained by looking at *janky* trades—those in which the number of shares traded is not a multiple of 10. We demonstrate evidence that janky trades are more common for high-frequency traders. Additionally, we suggest that a large number of small, janky trades may help signal that a large trade will happen shortly afterward.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Stock market trading

In the U.S. stock market, trades are made either by high-frequency traders (HFTs) or low-frequency traders (LFTs). High-frequency trades are often made through computers using proprietary algorithms. As a result, todays markets experience intense activity in the millisecond environment, where computer algorithms respond to each other at a pace 100 times faster than it would take for a human trader to blink (Hasbrouck and Saar, 2013). Due to computerized trading, investment companies can reduce the number of employees, helping the company reduce costs. Moreover, the declining costs of technology have led to its widespread adoption throughout financial industries (Hendershott et al., 2011). Because of that, high-frequency trades are more ubiquitous in this modern society. They make a significant portion of U.S. stock market trades; a high-frequency trader may make thousands or more stock trades in a day.

It is often thought that these trades are performed to take advantage of some inefficiency in the market. However, the aspects of the stock market that cause high frequency traders to act are poorly understood. We aim to gain understanding about the behavior of the high-frequency traders and to investigate what differences exist between high-frequency traders

and low-frequency traders.

Our analysis plan is as follows. We first perform statistical clustering on variables measuring the current status of the stock market using an unsupervised stock-trade dataset and investigate whether we see any separation between (potentially) high-frequency and low-frequency traders. Unsupervised here means that there is no variable in the dataset to identify high-frequency and low-frequency traders. We then use the NASDAQ-HFT dataset, which have the high-frequency and low-frequency trader identifiers, to investigate differences between these two types of traders. In particular, we uncover some interesting patterns of high frequency trading before and after certain large trades.

1.2 Clustering methods

Statistical clustering is the process of grouping units so that units within each unit are similar. Clustering can be helpful method for investigating the structure of data and to distinguish between groups in data (Hastie et al., 2009). statistical clustering also is useful in the design of experiments and observational studies (Higgins et al., 2015). We now give a brief overview of selected clustering methods.

1.2.1 K-means clustering

K-means clustering is one of the most widely used clustering methods. It is a partitional clustering technique that attempts to find a user-specified number of clusters K (Tan et al., 2013). K-means aims to minimize the sum of squares within each cluster:

$$\sum_{i=1}^{k} \sum_{x \in R_i} \|x - \mu_i\|_2^2$$

where μ_i is the center units of group R_i and $\|x - \mu_i\|_2$ is Euclidean distance between unit x and μ_i (Hastie et al., 2009). Solving the K-means clustering problem exactly is NP-

hard (Arthur and Vassilvitskii, 2007), and an exact solution can be found in $O(n^{dK+1})$ time, where n is the number of points to be clustered and d is the dimension of the covariate space (Inaba et al., 1994). However, the commonly used K-means algorithm provides a heuristic (often suboptimal) solution for this minimization problem in $O(n^2)$ time.

The K-means algorithm works as follows. An initial set of K centroids are selected. Each point is then assigned to the closest centroid, forming a cluster; there will be one cluster for each centroid. The centroid of each formed cluster is then computed, and clusters are updated by finding the points closest to the new centroid. This step repeats until all points are in clusters and the centroid of each cluster remains unchanged.

One of the reasons for the popularity of K-means clustering is that it obtains, usually, good results in a relatively quick amount of time. However, while K-means is relatively computationally inexpensive, it may still not work for a massive N due to its $O(n^2)$ run time. Hence, K-means alone may not be sufficient for a massive dataset.

K-means suffers from other problems as well. First, K-means often overfits individual points. That is, it may assign isolated data points to its own cluster. Second, the quality of the clustering may depend largely on the initial choice of centroids. Choosing initial centroids randomly is a common approach, but the result can be poor. Some methods have been developed to improve the quality of clustering by weakening the effect of initialization. Fränti et al. (1997, 1998) proposed genetic algorithms (GA) and Tabu Search (TS) that consider several possible initializations for K-means clustering. While these methods outperform choosing centroids at random, they also require significantly more computation. Fränti and Kivijärvi (2000) recently introduced randomized local search (RLS) that give similar results to GA and TS but runs much faster.

1.2.2 Hierarchical agglomerative clustering

Hierarchical clustering another method of cluster analysis in which, through building a hierarchy, a clustering with K clusters can be found for any arbitrary K. There are two main

types of hierarchical clustering: agglomerative and divisive. Agglomerative is a "bottom up" approach; it initially treats each unit as a cluster and then continues to merge two clusters together until only one cluster remains. Divisive is a "top down" approach; it assumes all units belong to one cluster initially and splits clusters until each cluster has only one observation inside. In this report, we focus on hierarchical agglomerative clustering (HAC).

HAC requires a dissimilarity measure d between every two units. Common dissimilarity measures include the Euclidean or Mahalanobis distance between the corresponding covariates between the units. Often, the dissimilarity measure will require to satisfy the triangle inequality: for any three units p, q, and r:

$$d(p,q) + d(q,r) \geq d(p,r) \tag{1.1}$$

HAC also requires a linkage criteria to measure the distance between two clusters and merge the least dissimilar clusters together. We discuss three different linkages. The first one is single linkage:

$$\min_{p \in R_i, q \in R_j} d\left(p,q\right)$$

The second one is complete linkage:

$$\max_{p \in R_i, q \in R_j} d\left(p,q\right)$$

The last one is average linkage:

$$\frac{1}{|Ri|\,|Rj|} \sum_{p \in R_i} \sum_{q \in Rj} d\left(p,q\right)$$

Where $|R_i|$ and $|R_j|$ respectively represent the size of the cluster R_i and R_j.

In single linkage hierarchical clustering, HAC merges in each step the two clusters whose two closest members have the smallest distance. In complete linkage hierarchical clustering, HAC merges in each step the two clusters with the smallest maximum pairwise distance.

Single linkage is good for with data with an non-elliptical shape. However, it is sensitive to outliers. Complete linkage is more robust to outliers but not works well when the covariate space is convex (Steinbach et al., 2004).

In practice, HAC does not tend to suffer from the same problems that K-means have of stopping at local minimum or subpar performance due to bad initialization of the algorithm. However, time complexity limits its application to massive data. In general, the complexity of agglomerative clustering is $O\left(N^2 \log\left(N\right)\right)$ (Rokach and Maimon, 2005).

1.2.3 Threshold clustering

Another recent development for clustering units is threshold clustering (TC) (Higgins et al., 2015). TC differs from the previous two methods in that it does not require the user to specify the number of clusters K in the clustering. Instead, the user specifies the minimum units t to be contained within a cluster.

Given a dissimilarity measure d that satisfies the triangle inequality (1.1), TC forms groups with at least t units and so that the maximum within-cluster dissimilarity—the maximum dissimilarity between two units in the same cluster—is small. That is, the objective of TC is to find a clustering $(R_1, R_2, \ldots, R_{K^*})$ that minimizes:

$$\max_{1 \leq i \leq K^*} \max_{p,q \in R_i} d(p,q)$$

Out of all such clusterings, the one obtained through TC has a maximum within-cluster dissimilarity at most a factor of 4 of optimal.

TC views units as a graph: units are vertices and an edge is drawn between every two vertices. Edges have weights equal to the dissimilarity of the corresponding units. The TC algorithm works as follows. First, a $(t-1)$–nearest–neighbors graph is constructed. Next, a set of units are selected as cluster seeds. The seeds satisfy two conditions; there is no path of two edges in the nearest–neighbors graph connecting two seeds and, for any non-

seed, there is a path of two edges or less connecting that non-seed to a cluster seed. Then, each seed grows a cluster comprised of that seed and any non-seed adjacent to the seed in the nearest–neighbors graph. Finally, any unassigned vertices are assigned to an adjacent cluster.

A significant advantage of TC is its computational cost. TC completes in $O(n)$ time and and space outside of the construction of a nearest–neighbors graph. Moreover, nearest–neighbor graphs are incredibly efficient to construct; most graphs can be constructed in $O(n \log n)$ time (Higgins et al., 2015). Hence, this clustering method is efficient enough for large datasets.

1.3 Hybridized threshold clustering

Hybridized threshold clustering (HTC) is a proposed method for clustering massive datasets. It works by applying TC repeatedly with a small t as a preprocess step, and after each iteration, reducing each TC cluster into one prototype point. Afterward, a more sophisticated clustering algorithm is applied to the prototype points. A final clustering of all points is performed by assigning each unit to the cluster to which its corresponding prototype unit belongs. Since TC is efficient enough for massive data, HTC can be a way of applying any given clustering algorithm to a massive dataset.

The HTC algorithm is as follows. First, TC with a pre-specified small t is applied the whole dataset to obtain clusters $R_1, R_2, \ldots, R_s$, where s denotes the number of clusters obtained through TC. Within each cluster, there is at least t units. Then, we reduce the units into a single point within each cluster, for example, by taking the centroid of the cluster. We repeat the previous two steps for r times until the dataset is sufficiently reduced. After that, we apply K-means clustering to the reduced dataset. Finally we back-out a cluster assignment of all units by assigning each unit to the cluster to which its corresponding prototype unit belongs.

1.4 The NASDAQ HFT dataset

	type	price	stock	Milis	BS	NoShares	Date	N
1	NN	104.10	AAPL	34200032	Buy	4	40109	1
2	HN	104.10	AAPL	34200033	Buy	396	40109	2
3	HN	104.10	AAPL	34200215	Buy	700	40109	3
4	NH	104.07	AAPL	34200419	Buy	100	40109	4
5	NH	104.08	AAPL	34200469	Sel	700	40109	5
6	HH	104.08	AAPL	34200483	Sel	400	40109	6
7	HN	104.10	AAPL	34200497	Sel	96	40109	7
8	NN	104.09	AAPL	34200914	Sel	116	40109	8
9	NN	104.09	AAPL	34200968	Buy	584	40109	9
10	NH	104.09	AAPL	34201103	Sel	20	40109	10

Figure 1.1: *A sample of units from the NASDAQ HFT dataset*

We look at two datasets for our analysis. The first dataset is the NASDAQ HFT dataset (Brogaard et al., 2014). The NASDAQ HFT dataset contains 8 variables in total: date, price, stock, BS, N, Milis, NoShares and type. It is comprised of 64,785,505 separate NASDAQ trades taken through 12 months of 2009. It includes 59 different stocks.

The variable type contains four values: HH, HN, NH and NN. H denotes high-frequency traders and N denotes a non-high-frequency traders (nHFT). The first letter is the type of trader who comes in and agrees to the limit order price. The second letter is the type of trader whose limit order is sitting there in the limit order book.

So "NN" means indicates that a nHFT demands liquidity and another nHFT supplies liquidity in a trade; "NH" is a nHFT agreeing to a limit order submitted by a HFT, as a nHFT demands and a HFT supplies; "HN" is HFT agreeing to a limit order submitted by a nHFT, as a HFT demands and a nHFT supplies liquidity; "HH" represents two HFT trading with each other, so both parties in the trade being HFTs Brogaard et al. (2014). For type "HH", there are 9,021,634 observations. For type "HN", there are 21,456,027 observations.

For type "NH", the total observations are 10,526,422. For the last type "NN", there are 23,781,422 observations.

"Milis" is the time in milliseconds. Trades are time stamped to the millisecond and identify the liquidity demander and supplier as a high-frequency trader (HFT) or non-high-frequency trader (nHFT). It starts from 34200000 which is 9 hours and 30 minutes. That means 9:30 am, the opening time of the stock market. The variable "BS" indicates if it is buy or sell. "NoShare" is the number of shares traded. Non-high-frequency traders we called it low-frequency traders as well. "Price" is the trading price of each trade.

1.5 The unsupervised stock trade dataset

	Spread	TotDept	nanos	Voly	NoShares	BS	DistPct
1	0.010891373	400	3.431542e+13	0.000000000	100	Sel	0.0002866151
2	0.010321101	400	3.432008e+13	0.016713268	500	Sel	0.0000000000
3	0.008893989	800	3.432008e+13	0.016713268	100	Sel	0.0000000000

SameDept	OppoDept	Type	ZVolu	N	Stock	Date
100	300	Lim	-0.2778883	1	ABCB	100316
100	300	Lim	-0.2778883	2	ABCB	100316
500	300	Lim	-0.2778883	3	ABCB	100316

Figure 1.2: *A sample of units from the unsupervised stock dataset*

The second dataset we consider is the unsupervised stock trade dataset. "Unsupervised" means that the dataset does not include identifiers as to whether the trade was performed by a high-frequency or low-frequency trader. This dataset includes 7,026,593 observations across one day, and contains 14 variables in total. It includes 300 different stocks. Figure 1.2 shows the unsupervised stock trade dataset.

Six of the variables are market characteristics. They are the most likely to effect the trader's decision to submit an order. The six market variables are "Spread", "TotDept", "Voly", "SameDep", "OppoDept" and "ZVolu". The variable "Spread" is the difference

between two prices. "TotDept" is the total depth on both side of the book at the time of the order (buy depth + sell depth). "SameDept" is the depth on the same side of the book as the order (buy depth when a buy order comes in, sell depth when a sell order comes in). "OppoDept" is the depth on the opposite side of the book as the order (buy depth for a sell order, sell depth for a buy order). The variable "Voly" is the volatility standardized; the number of standard deviations of a minutes average midpoint (of the spread) is away from the days average midpoint. The variable "ZVolu" is the volume standardized.

"BS" indicates if it is buy or sell. "Type" is either "Limit orders" or "Market orders". A limit order is an order to buy or sell a security at a specific price or better. A market order is an order to buy or sell a security immediately. "DistPct" is the distance of percent. "Type", "NoShares" and identifiers are variables describing the trade.

Chapter 2

Clustering the unsupervised stock trade dataset

To gain an understanding of the structure of trading behavior, we apply HTC to the unsupervised stock trades dataset. We cluster this dataset on all of the aforementioned market variables and aim to observe division within each cluster to better identify HFT and nHFT traders. Since this dataset is unsupervised, we do not know *a priori* which trades are performed by HFTs and which are performed by nHFTs.

We use HTC with a threshold $t = 2$ and 2 iterations of TC (we form prototype units twice). We then apply K-means with $K = 5$ clusters to form a clustering of all units. The choice of $t = 2$, $K = 5$, and 2 iterations of TC was selected after trying many different choices.

While HTC with K-means was able to be applied to this dataset, the inherent noise of the stock market data plus the few covariates we were able to work with made it difficult to extract much information from the clustering. We give some scatter plots as examples. Figure 2.1 is the scatter plot of the variables "Spread" and "TotDept" within cluster 1 for the first 30 minutes.

Figures 2.2, 2.3, 2.4 and 2.5 are the scatter plots of the same variables within cluster 2

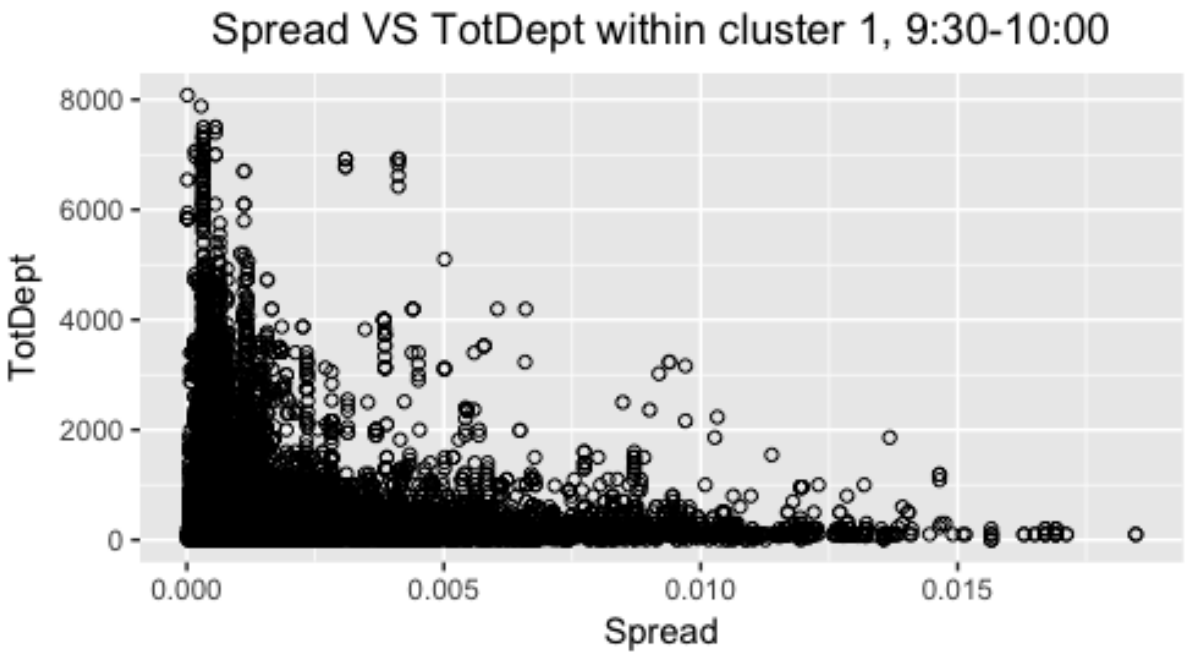

Figure 2.1: *A figure of scatter plot Spread VS TotDept within cluster 1 with time 9:30-10:00*

to 5 for the first 30 minutes. Some of the points form an obvious line while some of the points are concentrated in the lower right part of the figures. Within all five clusters, we can see that the two variables are correlated. These figures have similar patterns; most of the points are concentrated in the lower left corner and extending up and to the right. However, there is very little within-cluster behavior to distinguish between different types of trades. Scatter plots of other variables and other time periods show similar results, and hence, we omit them from this report. Hence, we were unable to glean valuable statistical inferences based on these clustering results.

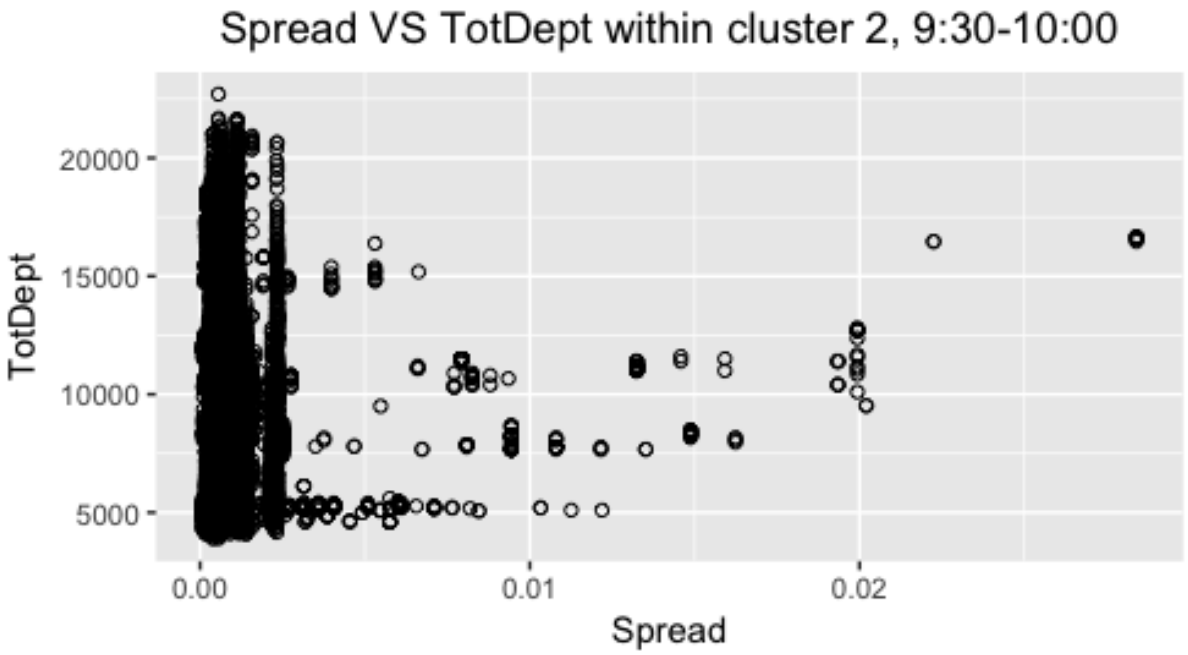

Figure 2.2: *A figure of scatter plot Spread VS TotDept within cluster 2 with time 9:30-10:00*

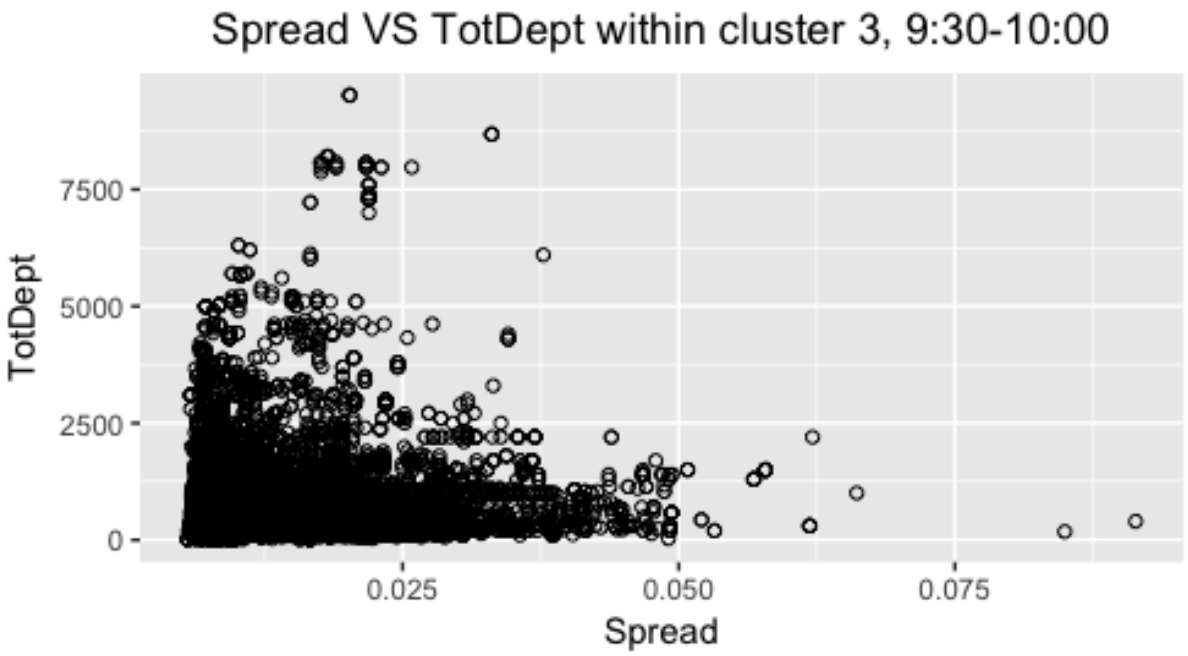

Figure 2.3: *A figure of scatter plot Spread VS TotDept within cluster 3 with time 9:30-10:00*

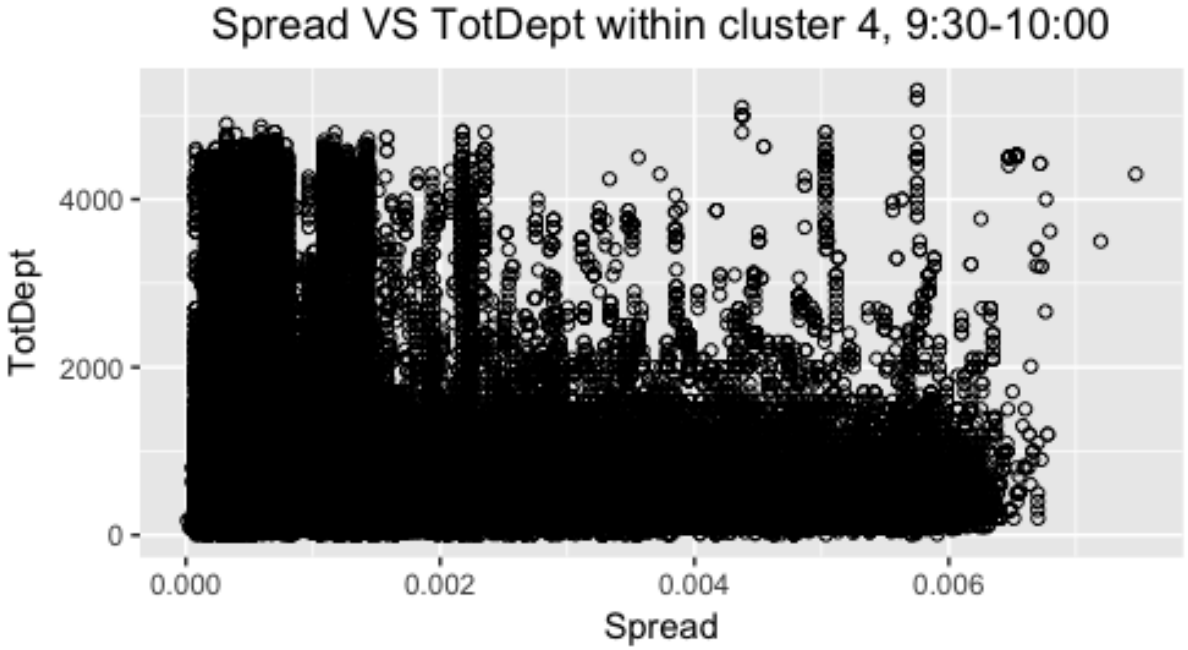

Figure 2.4: *A figure of scatter plot Spread VS TotDept within cluster 4 with time 9:30-10:00*

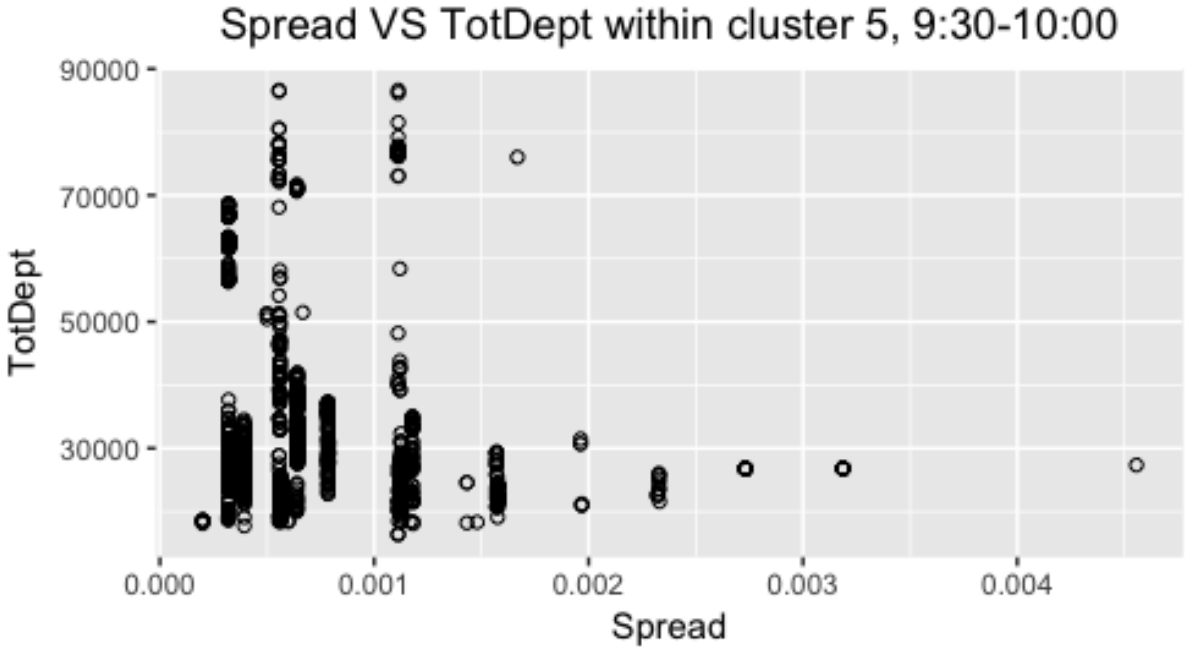

Figure 2.5: *A figure of scatter plot Spread VS TotDept within cluster 5 with time 9:30-10:00*

Chapter 3

Results from NASDAQ HFT dataset

3.1 Exploratory Data Analysis

We now turn our attention to the NASDAQ HFT dataset that has high-frequency non-high-frequency labels for each trade. We begin with an exploratory analysis to observe differences between the types of trades.

3.1.1 Analysis over entire day

We begin by looking at trade behavior across the entire trading day. We separate our analysis into small trades (those less than 100 shares), medium trades(those between 100 and 1000 shares), and large trades (those greater than 1000). The stock market starts trading at 9:30 am and closes at 4:00 pm, and we measure the activity of the stock market by looking at the number of trades d the number of shares traded at any given time. To reduce the amount of inherent noise in the data, we aggregate observations within groups across "time blocks" of 15 seconds. Thus, a trading day has a total of 1560 time blocks.

Figure 3.1 gives a plot of number of shares traded over time for the four types of traders. All four lines have a "U" shape. At 9:30 am and 4:00 pm, both low-frequency shares and high-frequency shares have sharp peaks. In particular, the number of low-frequency shares is

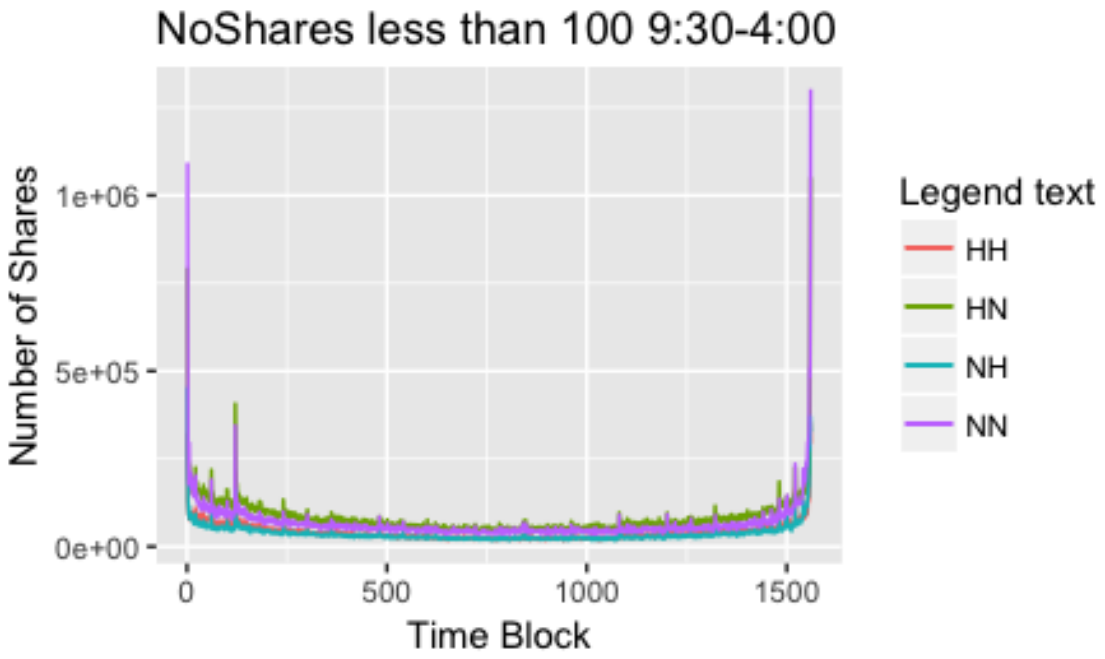

Figure 3.1: *A graph of number of shares from small trades VS Time from 9:30-4:00*

much higher than the high-frequency share when the stock market closes. After the decrease at the beginning, we notice that there are obvious peaks for all lines. Then they tend to decline gently and rise again. For most of the time, shares from HFTs line up with shares from nHFTs. Figure 3.2 plots the number of trades over time. The pattern is the same as the number of shares graph graph. But at the end of stock market, "HN" has the largest number of trades.

Figures 3.3 and 3.4 give the same plots for medium trades. The trend is similar to the previous plots for all four lines. However, it is worth noting that, for high-frequency shares, there are not high peaks at the beginning compared to low-frequency shares line. In Figure 3.3, the "NN" and"HN" lines basically overlap, as do do "HH" and "NH" lines. The gap between "NN" and "HH" lines is larger than the small shares figure. At the end of the figure, the low-frequency shares line is much more higher than the high-frequency shares line. Figure 3.4 shows similar patterns, but the "NN" line is above other three lines for starting and ending time.

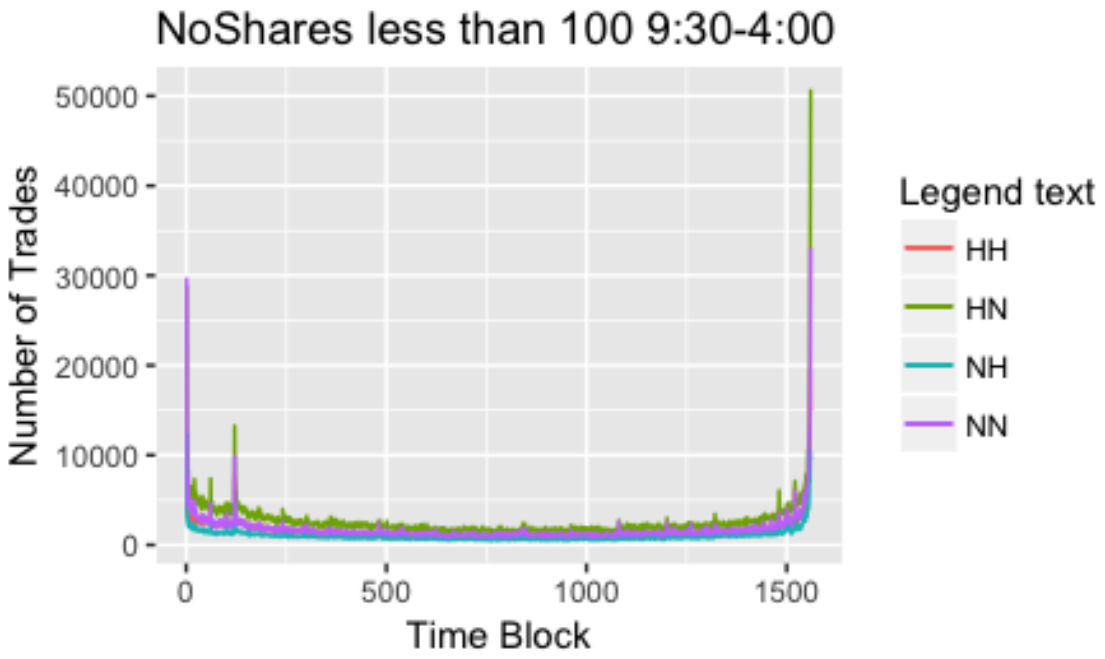

Figure 3.2: *A graph of number of trades from small trades VS Time from 9:30-4:00*

Figures 3.5 and 3.6 give the same plots for large trades. For figure 3.5, it is interesting that all lines do not have high peaks at the beginning. We can see that the "NN" line is above other three lines for the whole time period. Especially for the close time, there is dramatically high end as in previous graphs. Comparing to the "NN" line, the other three lines are more steady. They only increase a small amount at the end. However, for figure 3.6, at the beginning, "HN" line trades more than the other types and the peak is higher than the number of shares figure.

As we aggregate the three figures together, we will see the differences among different types of trades. As figure 3.7 shows, at the end of the market day, shares traded by "NN" for medium and large trades are far higher than the other types. It has a much larger range as well. Conversely, the other three lines don't have many fluctuations. From figure 3.8, we can see that most trades are medium-sized.

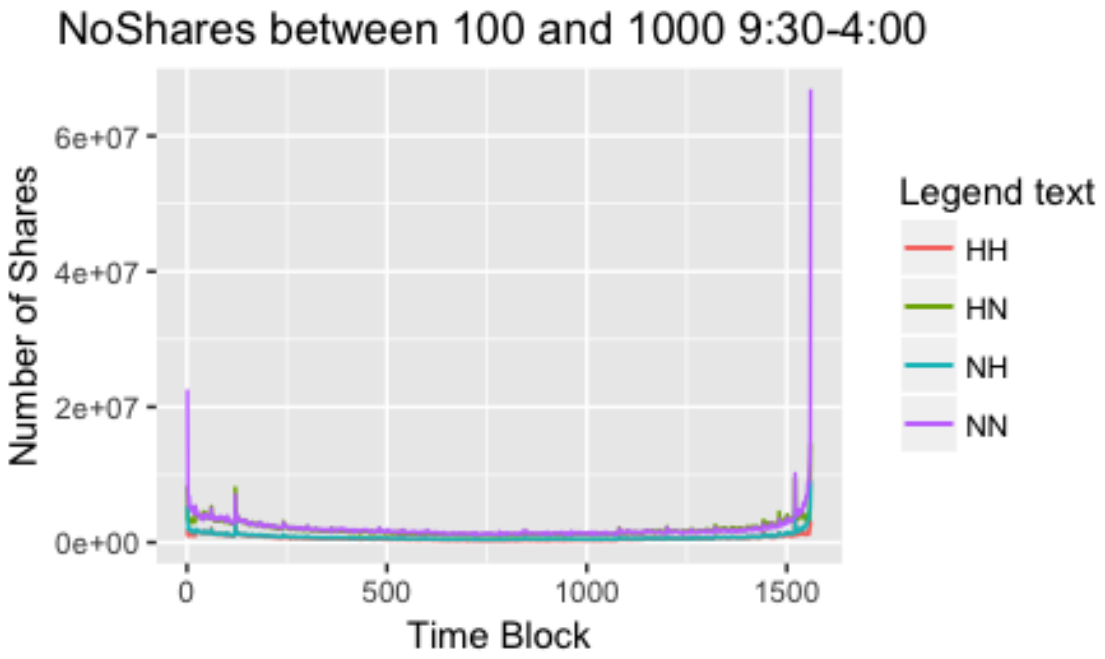

Figure 3.3: *A figure of number of shares from medium trades VS Time from 9:30-4:00*

3.1.2 Beginning, middle, and ending of day behavior

To better illustrate differences between the types of trades, we isolate these plots to three time periods: 9:30 am - 10:30 am, 11:30 am -12:30 pm and 3:30 pm - 4:00pm. The reason why we only focus on these specific time periods is because there are more high peaks and fluctuations during these time periods, especially when the market starts or is about to end, allowing us to better view differences in the high-frequency and low-frequency traders. The rest of Section 3.1 will focus on the plots and interpretation for these time periods.

Since we specified the type of each trade as "HH", "HN", "NH" and "NN", we make more plots based on different types of trades.

Figure 3.9 gives the number of shares traded for the four types of trades for the first hour of trading. We have 120 time blocks and each block is 15 seconds. For all four lines, the number of shares traded are all high at the beginning. The highest one is the "NN" line. The second highest one is the "HN" line. The third one is "NH" line. The last one is "HH" line. After the beginning time, all lines decrease and they have similar fluctuations. We can

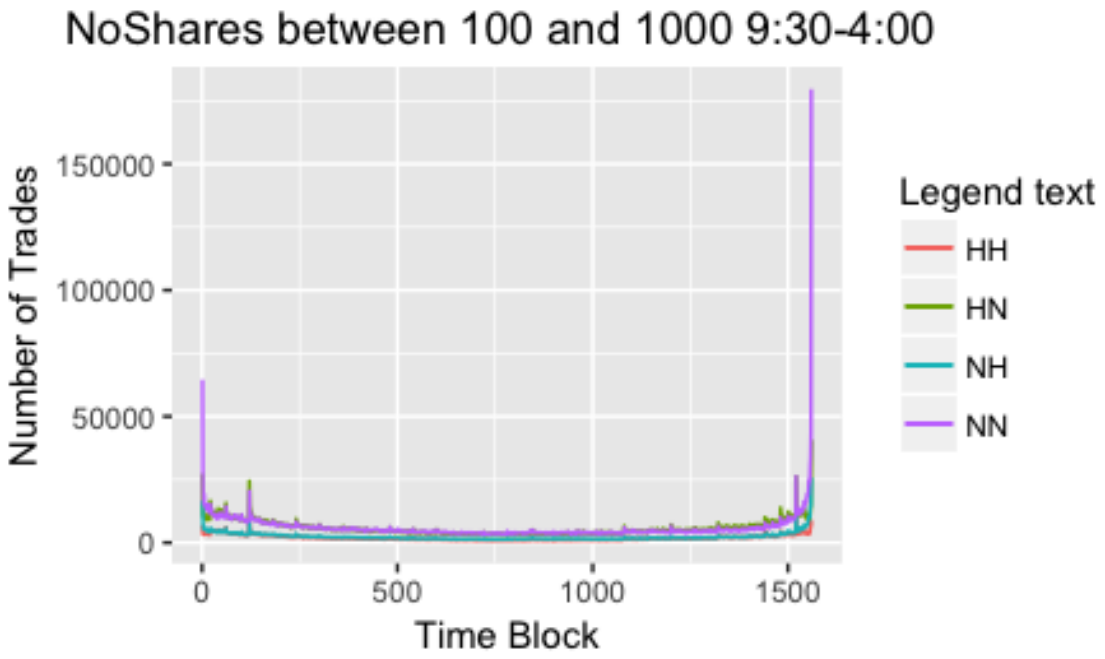

Figure 3.4: *A figure of number of trades from medium trades VS Time from 9:30-4:00*

see that starting from around time block 23, the four lines are more distinguishable. "HN" line is above all other lines and "NH" line is the lowest one. There are obvious high peaks around time block 120 which is 10:00 am for both shares and trades graphs. For number of trades graph, Figure 3.10, "HN" is above the other three lines except for the beginning time.

Figure 3.11 is the trades between 100 and 1000. Like the previous graph, the "NN," "NH," and "HN" lines have high peaks when stock market starts. However, for the "HH" line, the number of shares traded is not large. Before time block around 23, the four lines can be clearly distinguished. "NN" line is the top line and "HH" line is the bottom line. After that time block, we can see the "NN" and "HN" lines are hard to divide and so are the "NH" and "HH" lines. With similar fluctuation, "NN" and "HN" lines are above the other two lines. The number of trades graph, Figure 3.12, shows the same fluctuations and patterns as the number of shares graph. Much like for the small trades, high peaks show around 10:00 am for the medium-sized trades.

Figure 3.13 shows this same graph for trades greater than 1000 shares for the first 30

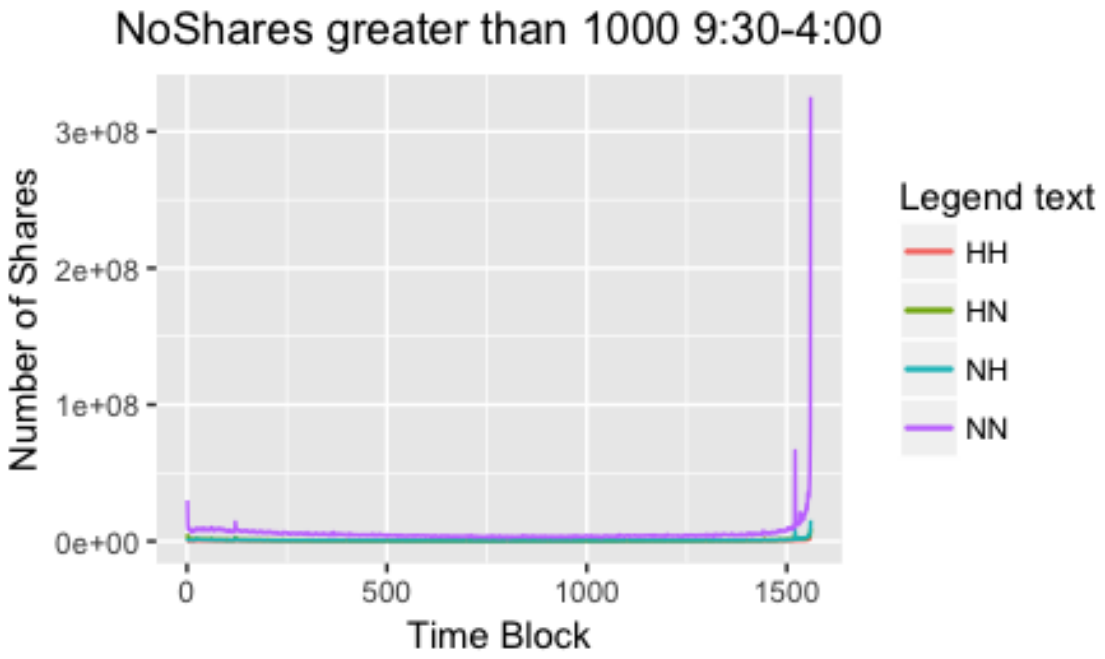

Figure 3.5: *A graph of number of shares from large trades VS time from 9:30-4:00*

minutes when stock market starts. While the "NN" line has a high initial peak, other three lines do not show this peak. We can see for the whole time period, "NN" line is obviously above all other three lines. Each line can be easily distinguished for both shares and trades graphs. "NN" and "HN" have more fluctuations than "NH" and "HH". The high peaks around 10:00 am are also significant except for "HH" type. Comparing with Figure 3.13, Figure 3.14 shows that the gap between "NN" and "HN" gets smaller.

Figure 3.15 plots the number of shares traded for trades less than 100 from 11:30 am to 12:30 pm. The lines except for the "HH" line have huge number of trades at 11:30 am. Besides that, they all have similar fluctuations. Around time block 30, there is an obvious peak for all three lines. However, the "HH" line is more stable and no visible peak. For "NN" and "HN", we can see that at the beginning, the "NN" line is over the "HN" line. But after the first few time blocks, the "HN" line is over the "NN" line. The same pattern shows in the number of trades graph in Figure 3.16.

Figure 3.17 gives the plot for trades between 100 and 1000. Both the "NN" and the

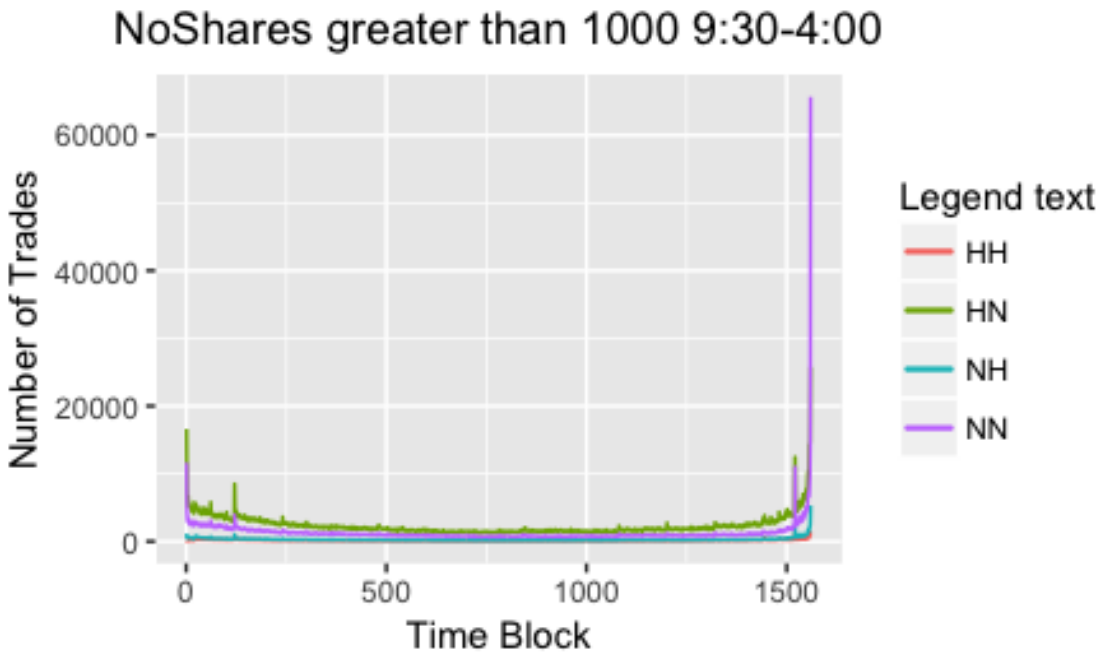

Figure 3.6: *A graph of number of trades from large trades VS time from 9:30-4:00*

"HN" lines fluctuate more than the other two lines. Comparing to "NN" and "HN", the "NH" and "HH" lines are more steady. The fluctuation range of the "HH" and "HN" lines is wider than the lower two lines. There is a large gap between the "HH" and "HN" lines and the "NH" and "NN" lines for both Figures 3.17 and 3.18. The trend of the four lines is slightly decreasing over this time period.

Figure 3.19 shows the trades greater than 1000 for time period 11:30 am to 12:30 pm. For Figures 3.19 and 3.20, each line is clearly distinguishable. The "NN" line is on top. The second highest one is the "HN" line. The third one is "NH" line. The last one is "HH" line. Both the "NN" and "HN" lines show similar fluctuations. However, for the "NH" and "NN" lines, both of them are more stable and do not fluctuate too much.

Figure 3.21 gives the line plots for trades smaller than 100 when the stock market closes within the last 30 minutes. All four lines are close to each other and they all have peaks at same time blocks such as around 40, 80 and 100. We can see that most of the time, the "HN" line is over the "NN" line. But at the end of the time period, the "NN" line increases

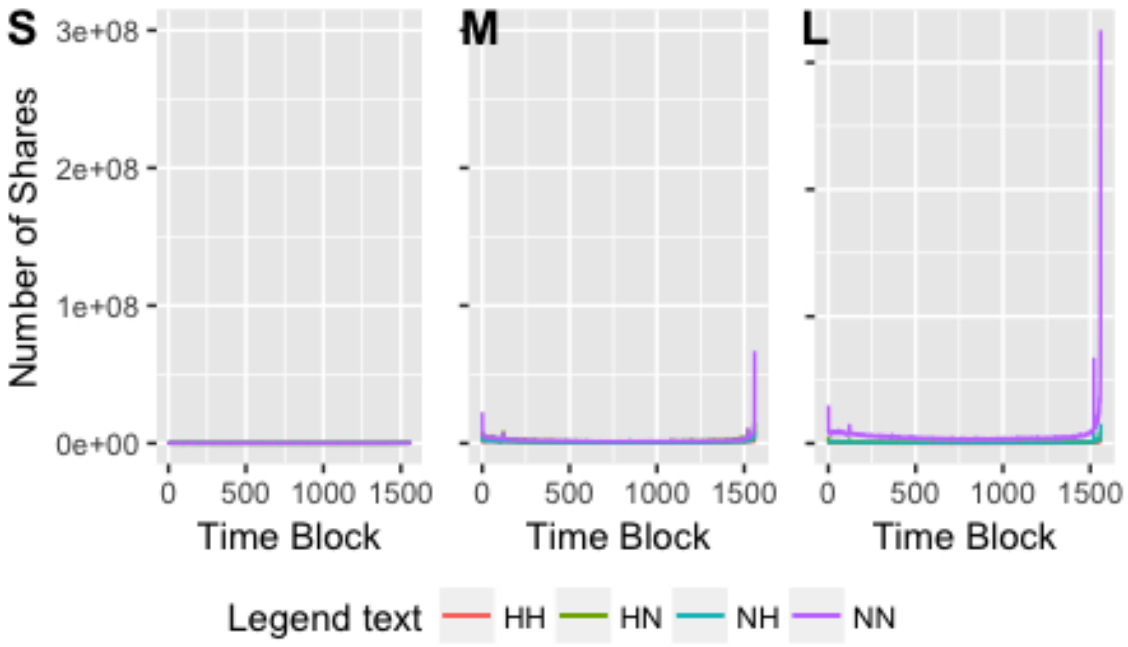

Figure 3.7: *A comprehensive shares graph for time 9:30-4:00*

rapidly. All four lines increase when stock market ends but "HH" and"NH" lines do not have as many shares traded as the "NN" and "HN" lines do. The difference between Figures 3.21 and 3.22 is that, at the end of the stock market, "NN" has the largest number of shares but "HN" has the largest number of trades. Although "HH" has the least number of shares, it does not have the smallest number of trades.

Figure 3.23 shows this plot for medium-sized trades. All four lines do not show too much fluctuation until time point 75. We can see there is not much trading going on before the end. It is hard to distinguish between "NN" line and "HN" line, and it is hard to distinguish between the "NH" line and "NN" line. After time point 75, all lines except for the "HH" line increase. In particular, the "NN" shows significant growth. "HH" line only increases a little and it shows a decreasing trend by the end.

Figure 3.25 gives this graph for large trades. The trends for each line are similar to the previous graphs. However, the four lines are very close to each other before the peaks appear at point around 80. The "NN" and "HN" lines have much higher peaks than the "HN" and

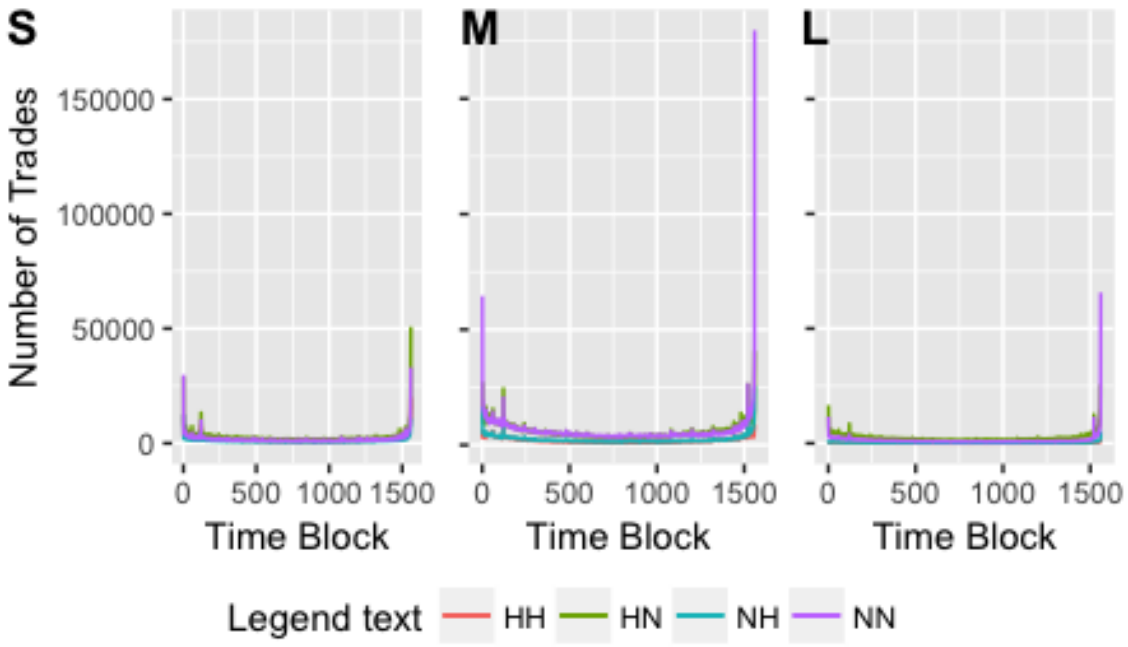

Figure 3.8: *A comprehensive trades graph for time 9:30-4:00*

"HH" lines. After the high peaks, the "NN" line starts to increase; by the end of the time period, the increase in this line is much larger than other lines. The "HN" and "NH" lines also increase at the end, but not much. The "HH" line does not show an increase. It is stable and smooth for the whole time period.

The number of trades figures fluctuate the same as the number of shares figures for both medium and large trades size.

Figure 3.27, 3.28 and 3.29 aggregate the previous plots together based on small trades, medium trades and large trades within the three time periods. We set the same trades range for all three figures to see the difference among different size trades.

In general, low-frequency trades tend to be larger than high-frequency trades. For every graph, the "HH" and "HN" lines are always higher than the other two lines. They have a huge number of shares traded. We can see that for the first 30 minutes, there are not many differences between low-frequency and high-frequency trades for small trades. However, for medium trades and large trades, the number of shares traded by low-frequency traders is

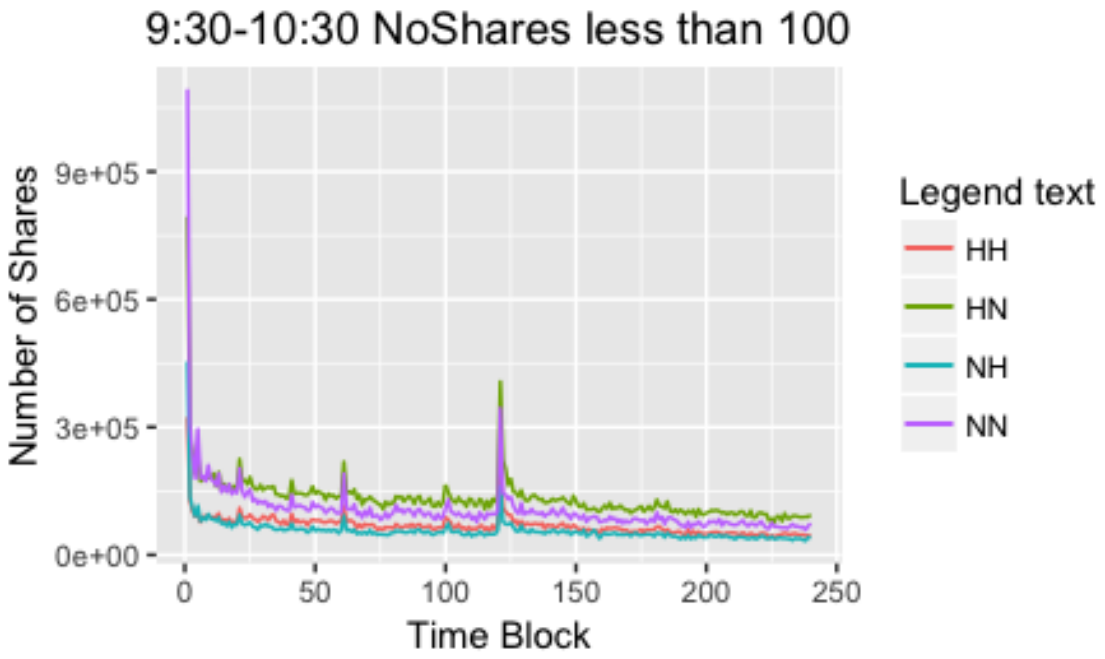

Figure 3.9: *A graph of number of shares from small trades VS Time 9:30-10:30*

larger. Especially for large trades, low-frequency traders trade many more shares than high-frequency traders. However, for the middle one hour, for medium trades and large trades, we can see some gaps between different types of trades. For large trades, the gap is not as huge as medium one. For the last 30 minutes, it is hard to see the difference among these lines except at the end of large trades part.

Low-frequency trades tend to have higher peaks. It is more obvious that low-frequency trades have much higher peaks at the beginning for both medium and large trades. Even for small trades, the low-frequency trades peak is higher than high-frequency ones.

If we look at all the graphs together, we can notice that all four types of the trades have almost the same fluctuations. When low-frequency trades have high peaks, high-frequency trades also have high peaks. When stock market stars or closes, the low-frequency trades line tends to have drastic increases for all three trades sizes. For the last 30 minutes, there are four apparent peaks.

Focusing on figure 3.30, 3.31 and 3.32, we can conclude that large trades have large

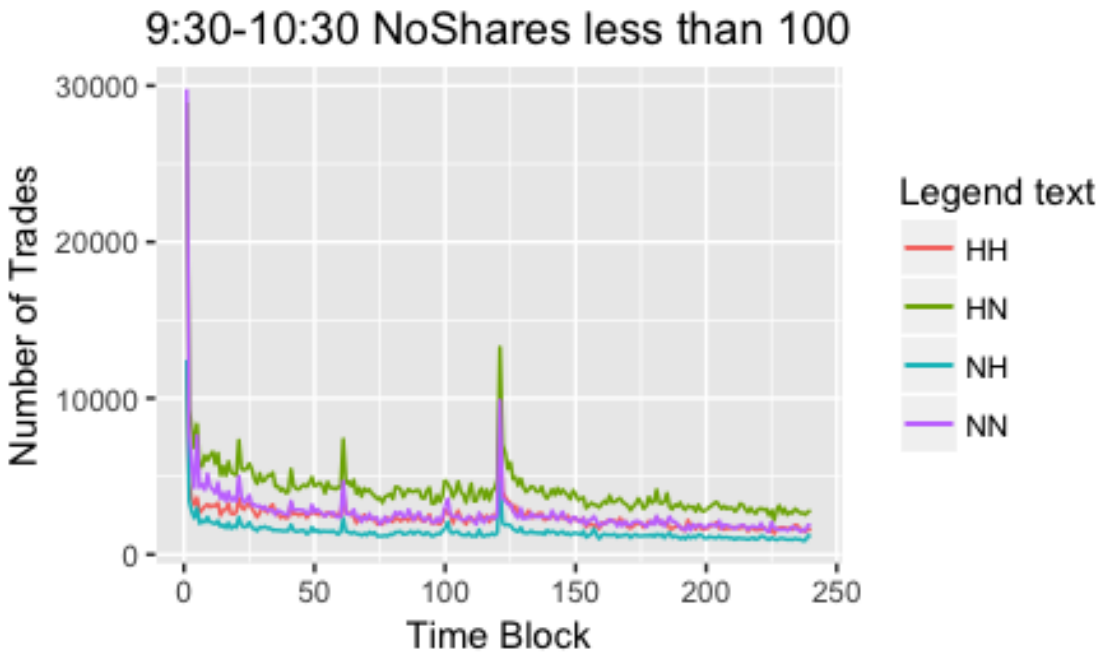

Figure 3.10: *A graph of number of trades from small trades VS Time 9:30-10:30*

number of shares but most trades are medium-sized.

3.2 Behavior before and after large trades

3.2.1 Proportions of janky trades

We investigate whether, before large trades, if there are a large number of small janky trades. Janky trades are defined as trades of less than 100 shares. We suggest that a large number of small, janky trades may help signal that a large trade will happen shortly afterward. For our report, we define trades of 100,000 or more shares as large trades. We isolate our analysis to only the 200 trades before and 100 trades after a large trade (including the large trade).

Table 3.1 shows the proportion of janky trades by type. The second column are the percentages of small trades which by different types of trade. The third column are the percentages of janky trades overall. In general, we see that small janky trades are more common for high-frequency traders.

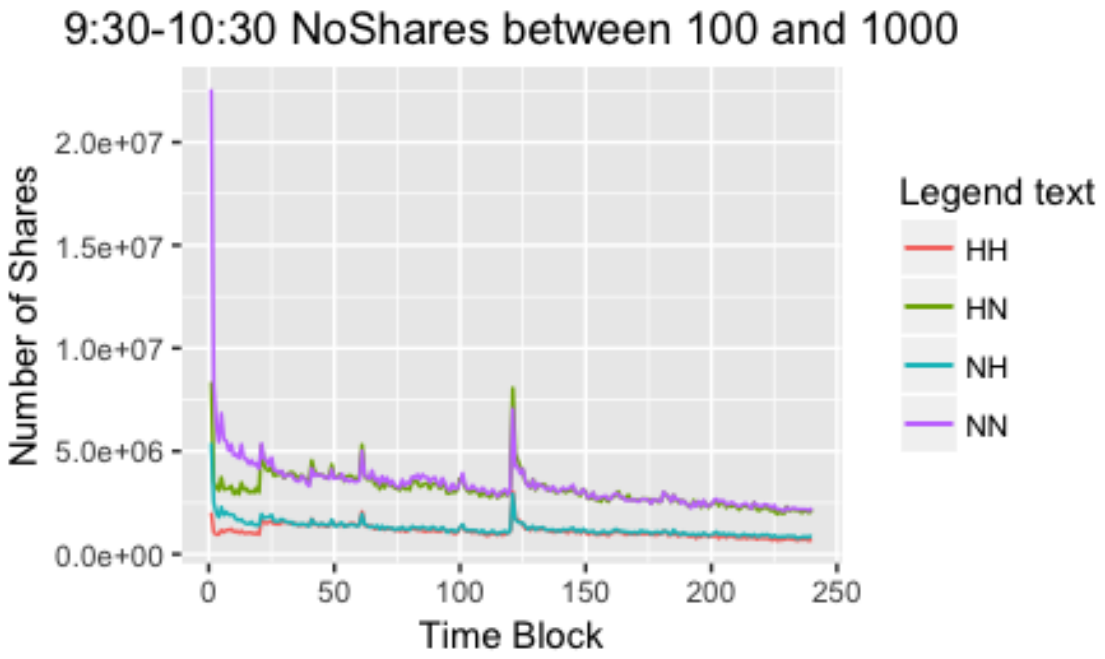

Figure 3.11: *A graph of number of shares from medium trades VS Time 9:30-10:30*

Types	% of janky trades before big trades	% of janky trades overall
HH	30.77%	5.82%
HN	27.50%	10.06%
NH	20.98%	3.85%
NN	20.76%	6.67%

Table 3.1: *Proportions of janky trades by type*

Table 3.2 shows the large trades percentage by type. We can see that no trade of 100,000 or more shares is performed by "HH". In fact, there are only 2.1% that have a HFT trader. Therefore, we conclude that the large trades are mostly done by non-high-frequency traders.

3.2.2 Less than 100 size trades before and after large trades

We try to exploring more about the behavior of these large trades. Considering there there appears to be many trades of fewer than 100 shares around large trades, we make a bar plot comparing the proportions of those small trades before and after large trades based on different types.

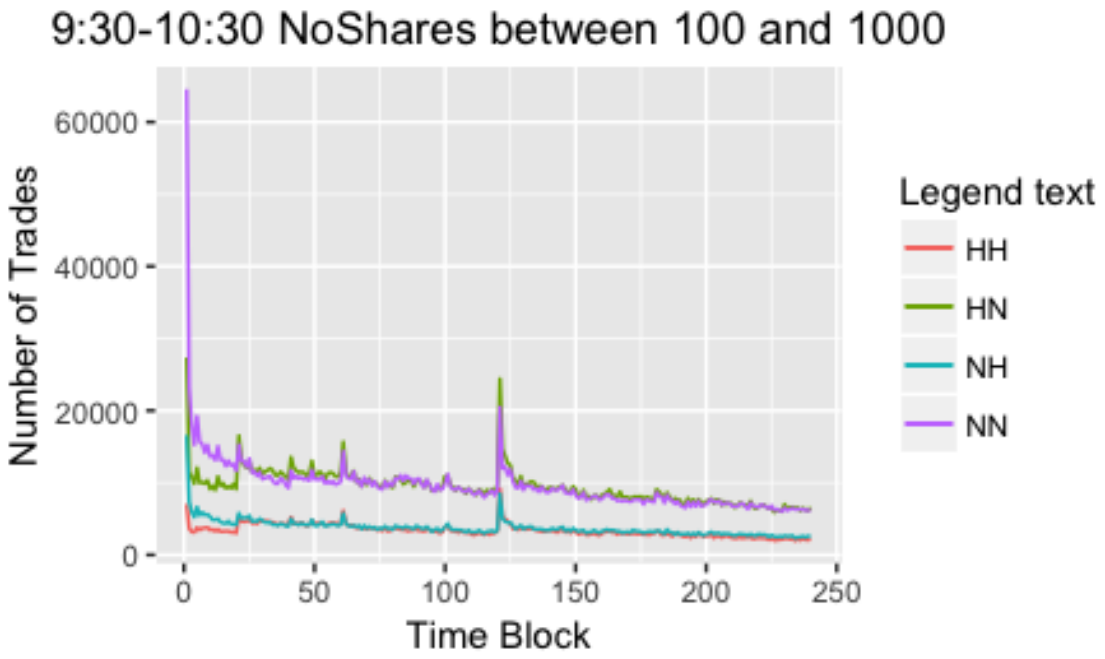

Figure 3.12: *A graph of number of trades from medium trades VS Time 9:30-10:30*

Type	% of large trades
HH	0 %
HN	1.8%
NH	0.3%
NN	97.9%

Table 3.2: *Proportions of each type of large trades*

As Figure 3.33 shows, for both before and after, the "HH" and "HN" proportions are very close. Focusing on "NH" type, the proportion of "NH" trades of 100 or fewer shares before a large trade slightly smaller than "HH" and "HN" while for after, it is slightly large. However, the proportion of "NN" trades of 100 or fewer shares is much smaller before a big trade, and after a big trade, it is much larger. There are big differences between non-high-frequency trades and high-frequency-trades.

After removing all the 100 size trades, as Figure 3.34 shows, we can see that the proportions of type "NN" doesn't change much before and after a large trade. For "HH" and "NH", the proportions decrease; for "NH" this proportion decreases substantially.

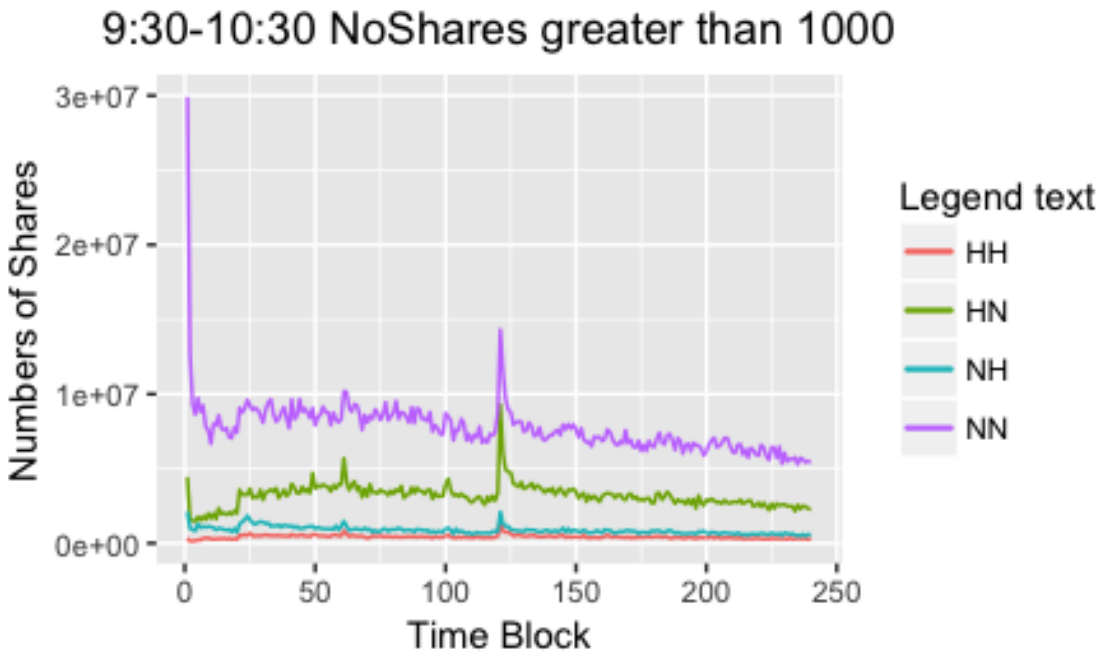

Figure 3.13: *A graph of number of shares from large trades VS Time 9:30-10:30*

Figure 3.35 gives an example showing the behavior of a relevantly large trade. It is clear that there are a large amount of janky trades before the trade of 13,500 shares. The large trade is performed by "NN" while all the small trades before this trade are performed by "NH." If we notice the prices of these trades, the large trade hits the lower price than all the small trades. Furthermore, all the small trades are non-high-frequency traders selling to high-frequency traders.

Figure 3.36 shows another example of this behavior. The same trend appears in this example. After all the small trades which are done by "NH", there is a large trade of 9,300 shares and a larger trade of 118,800 shares, both of which are performed by "NN."

As figure 3.37 shows, there are many trades after the large trade of 13,500 shares performed by "HN." Recall, before the large trade, all the small trades were performed by "NH." That means the nHFTs are selling and the HFTs are buying. That is, that high-frequency traders are buying orders both before and after that large trade.

As figure 3.38 shows, there is a trade of 11,145 shares performed by "NN" as well. After

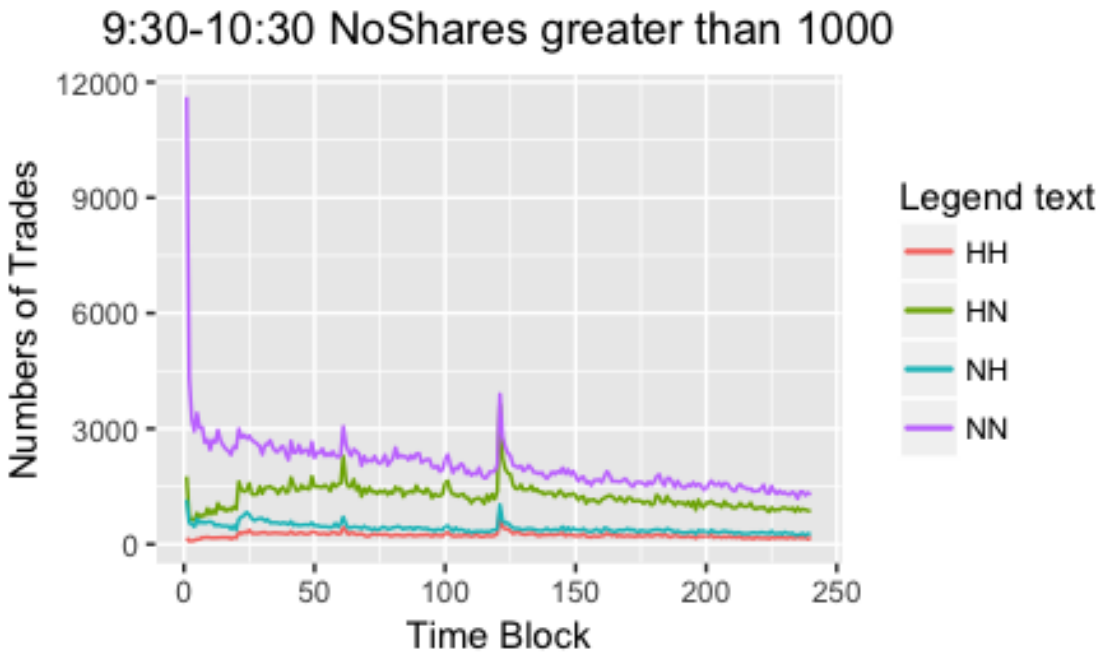

Figure 3.14: *A graph of number of trades from large trades VS Time 9:30-10:30*

that, there are some trades performed by "HH"; One sells 1,677 shares and two buy a small amount of shares. Afterward, there are trades performed by "NH" and "HN". For these trades, high-frequency traders are selling hundreds-to-thousands of shares.

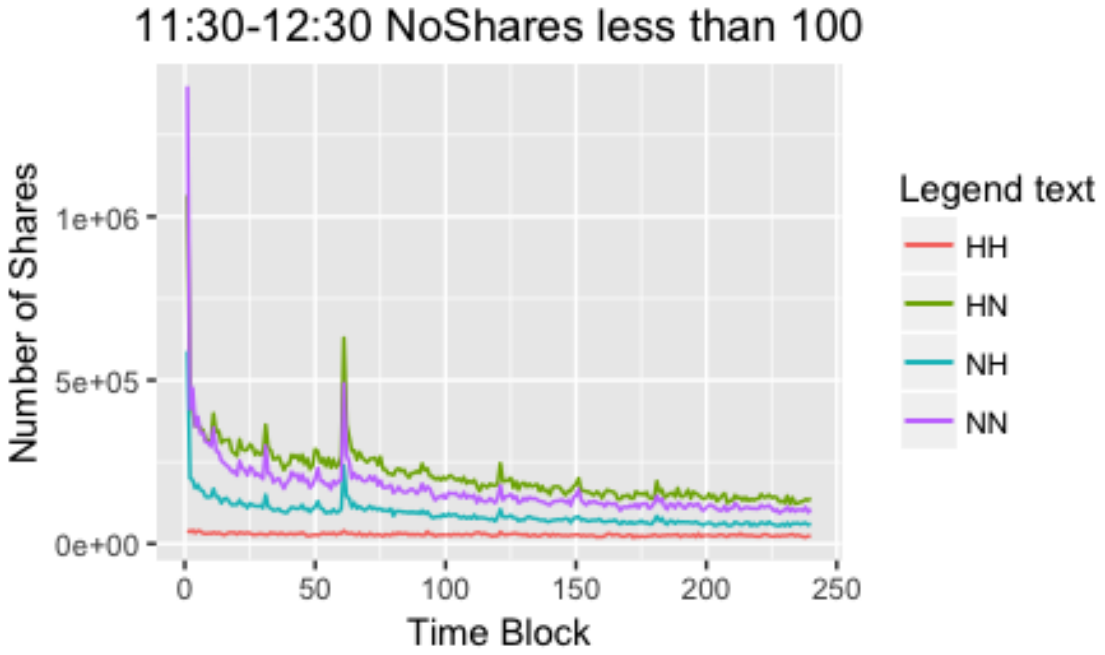

Figure 3.15: *A graph of number of shares from small trades VS Time 11:30-12:30*

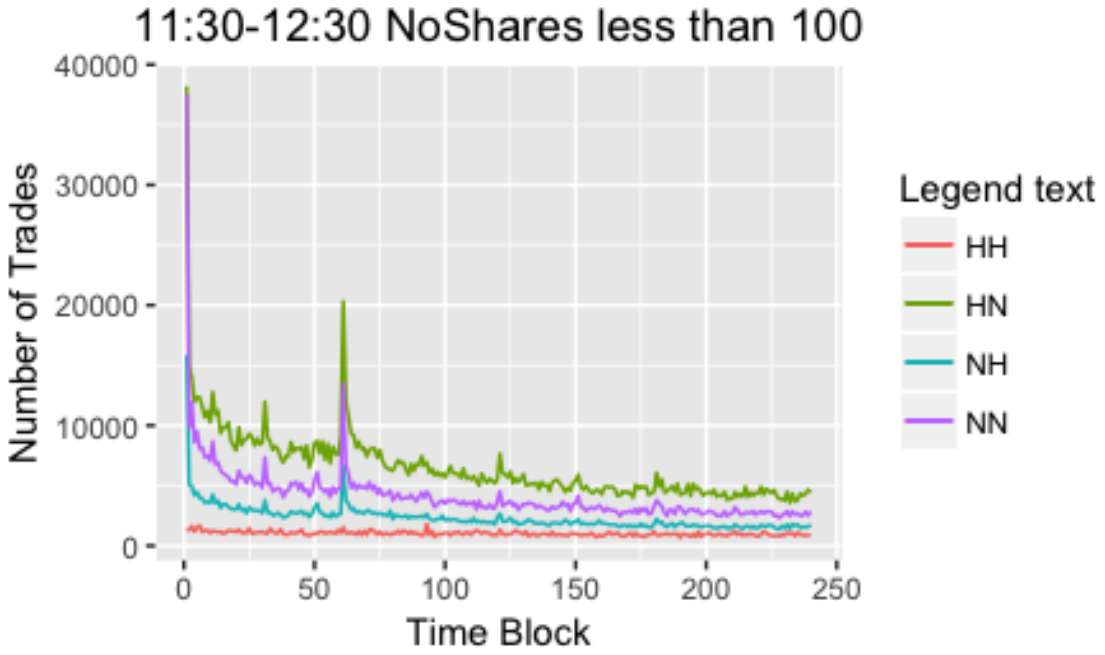

Figure 3.16: *A graph of number of trades from small trades VS Time 11:30-12:30*

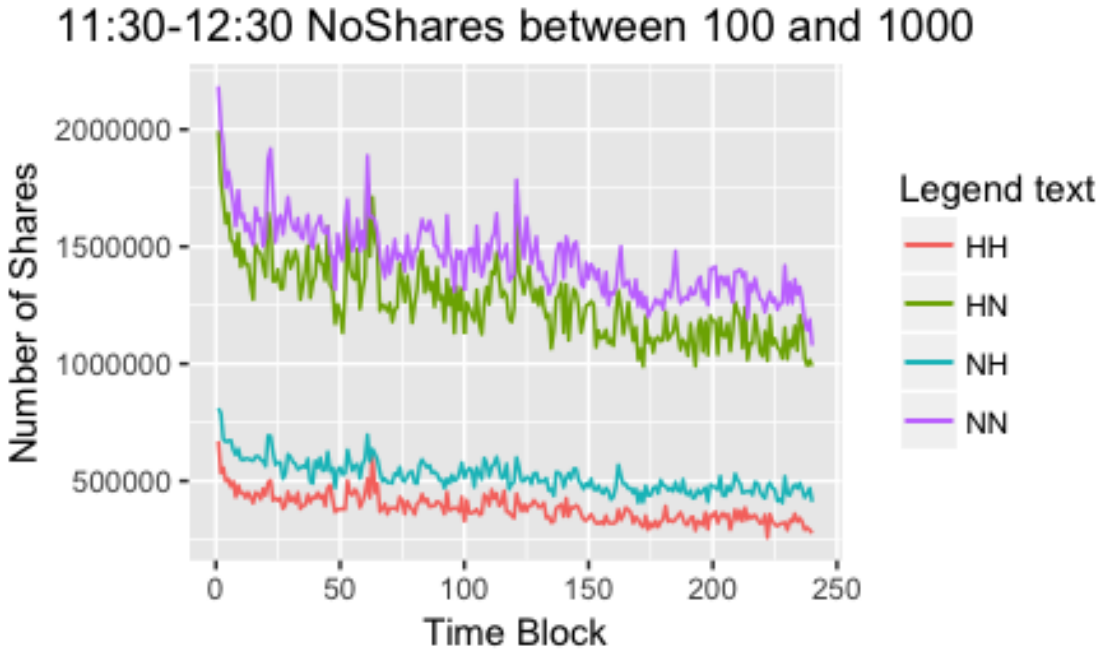

Figure 3.17: *A graph of number of shares from medium trades VS Time 11:30-12:30*

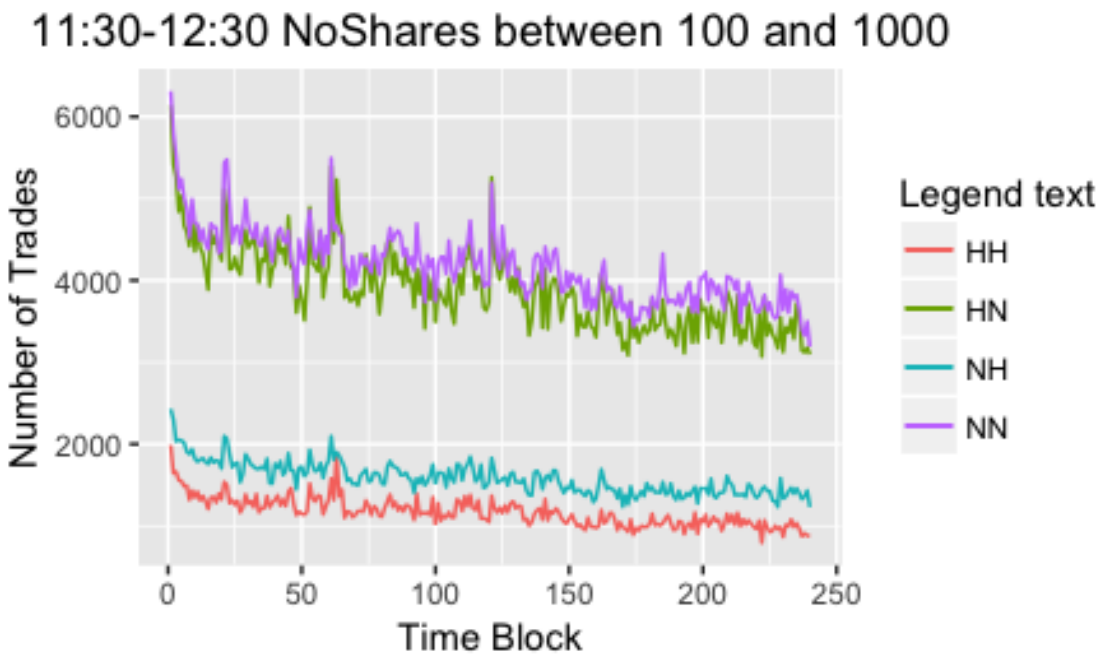

Figure 3.18: *A graph of number of trades from medium trades VS Time 11:30-12:30*

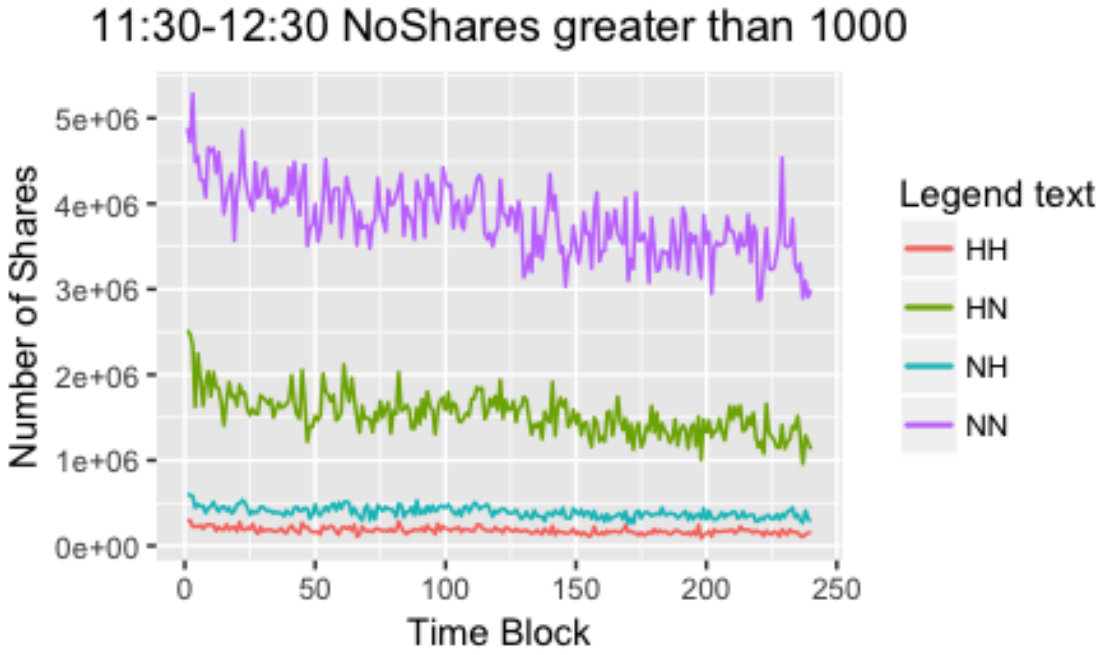

Figure 3.19: *A graph of number of shares from large trades VS Time 11:30-12:30*

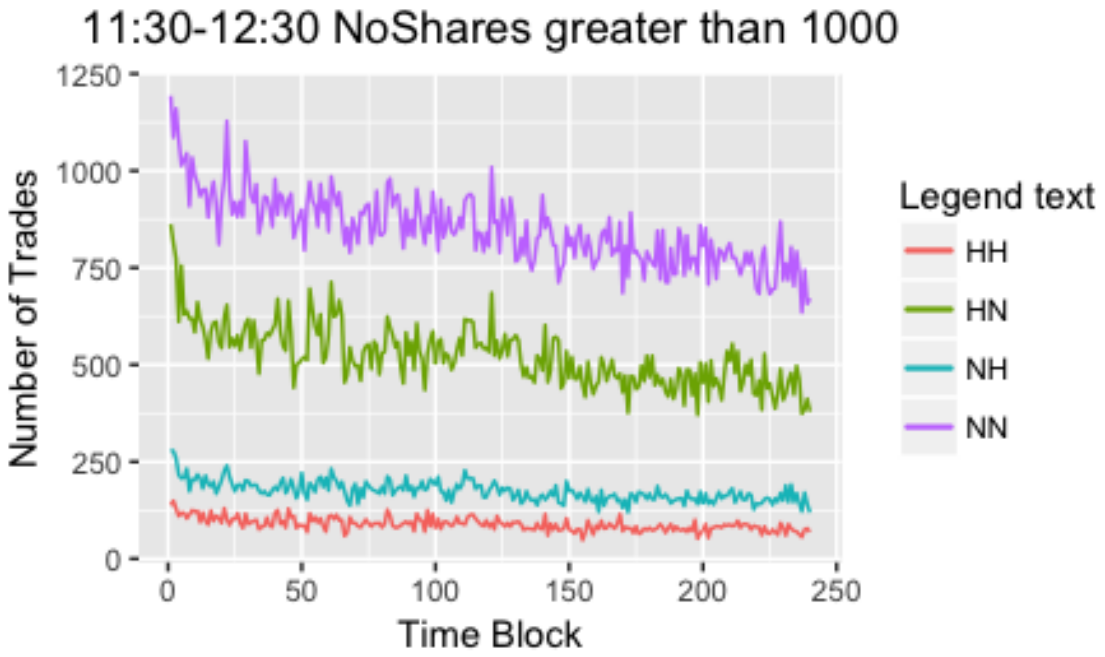

Figure 3.20: *A graph of number of trades from large trades VS Time 11:30-12:30*

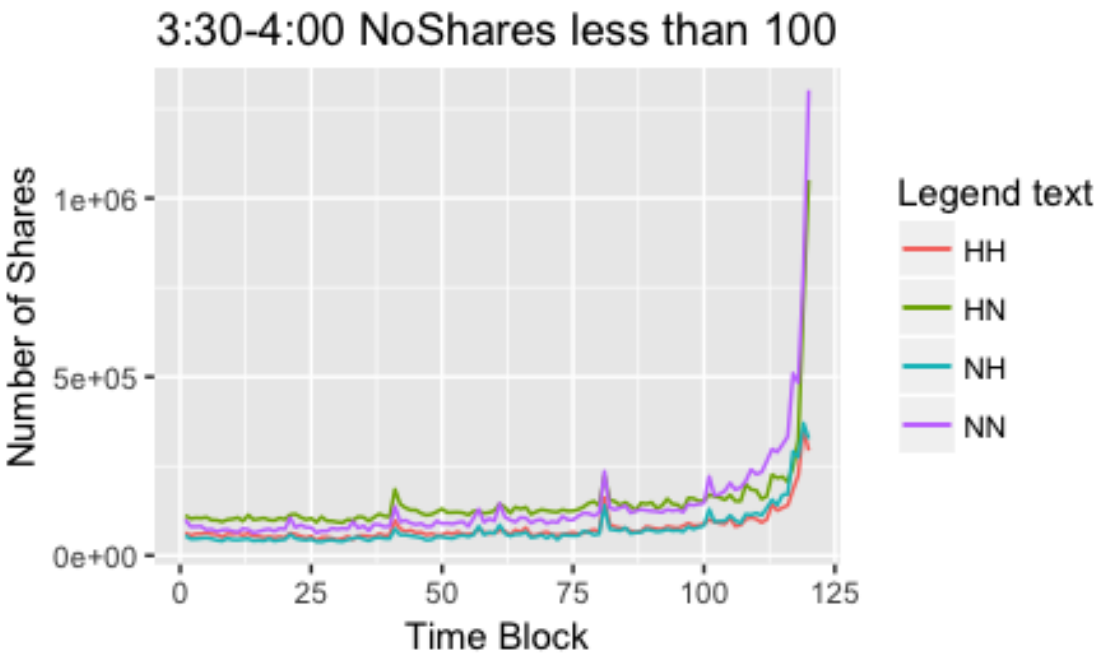

Figure 3.21: *A graph of number of shares from small trades VS Time 3:30-4:00*

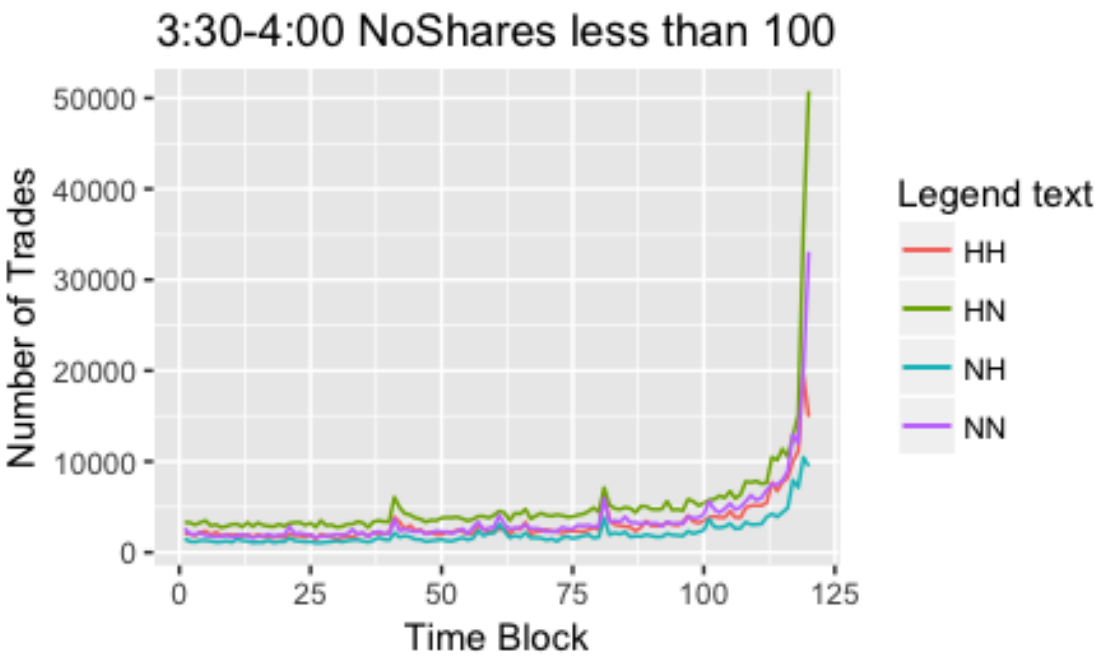

Figure 3.22: *A graph of number of trades from small trades VS Time 3:30-4:00*

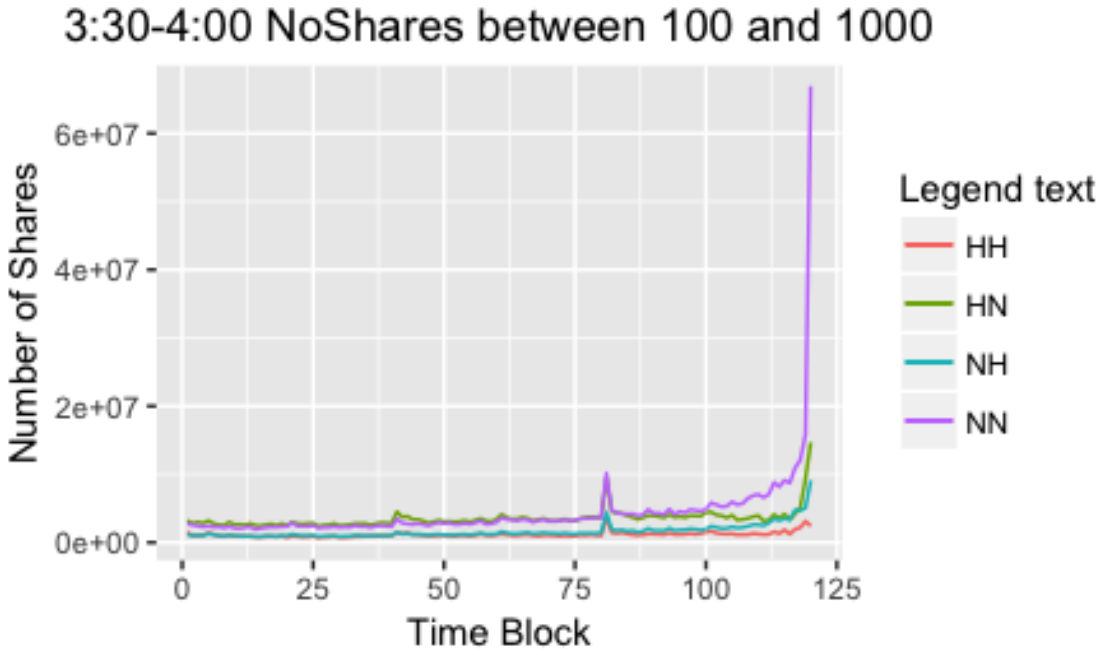

Figure 3.23: *A graph of number of shares from medium trades VS Time 3:30-4:00*

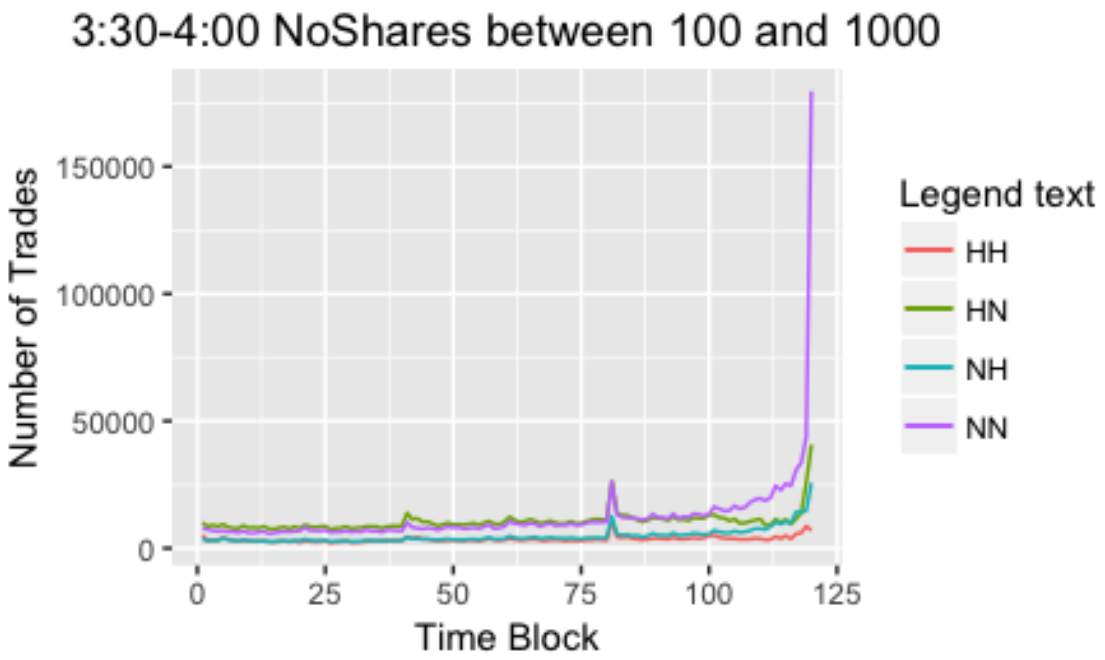

Figure 3.24: *A graph of number of trades from medium trades VS Time 3:30-4:00*

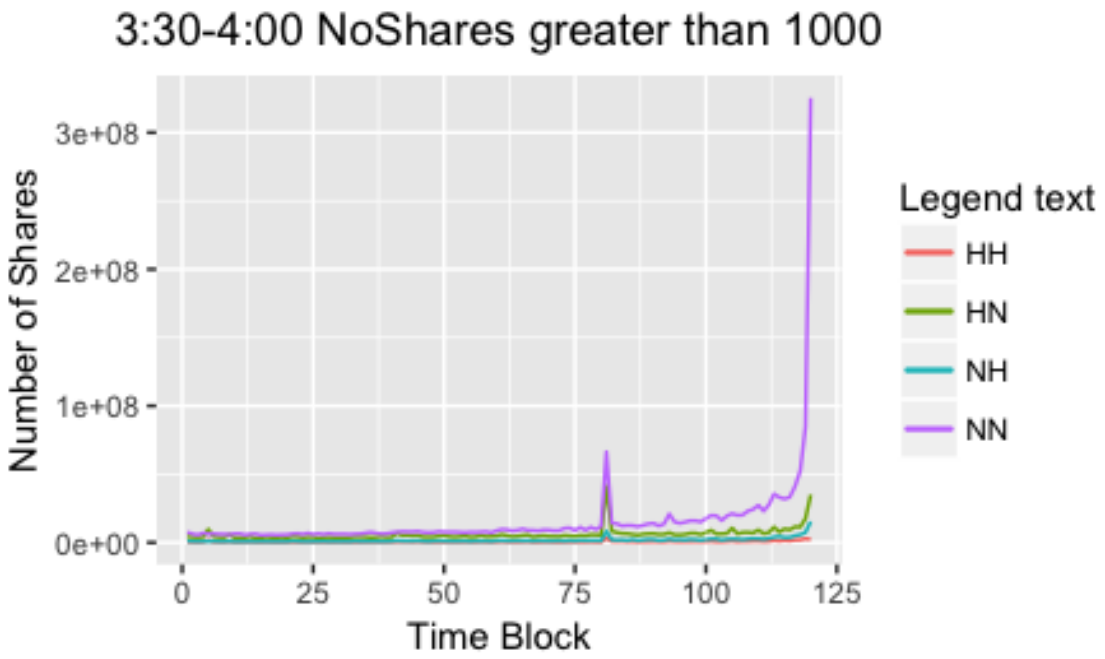

Figure 3.25: *A graph of number of shares from large trades VS Time 3:30-4:00*

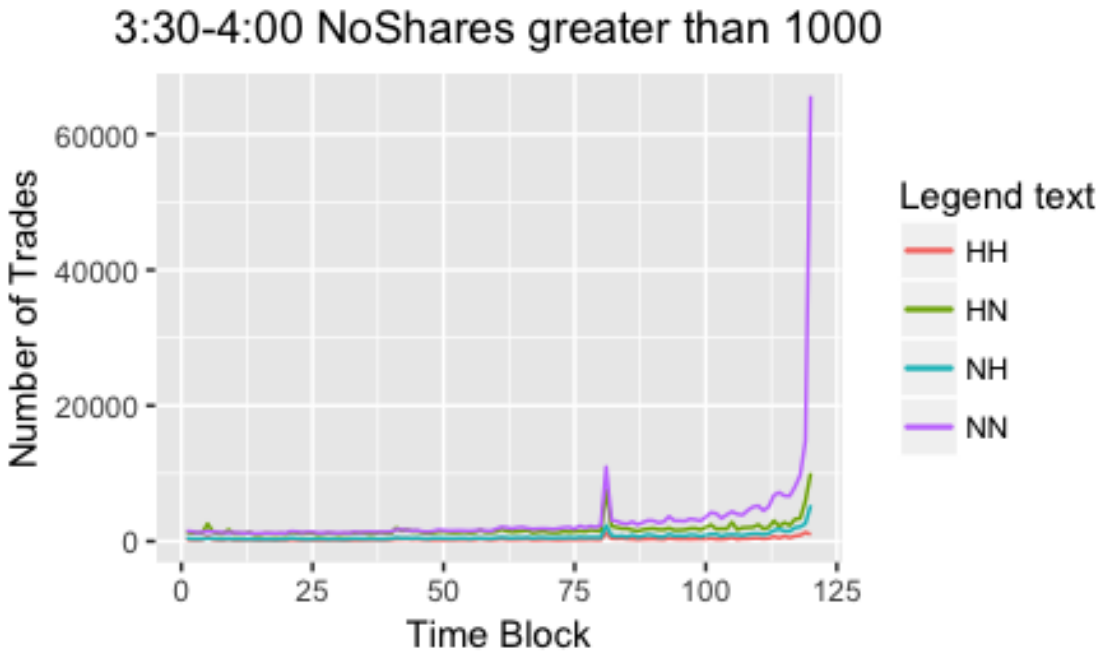

Figure 3.26: *A graph of number of trades from large trades VS Time 3:30-4:00*

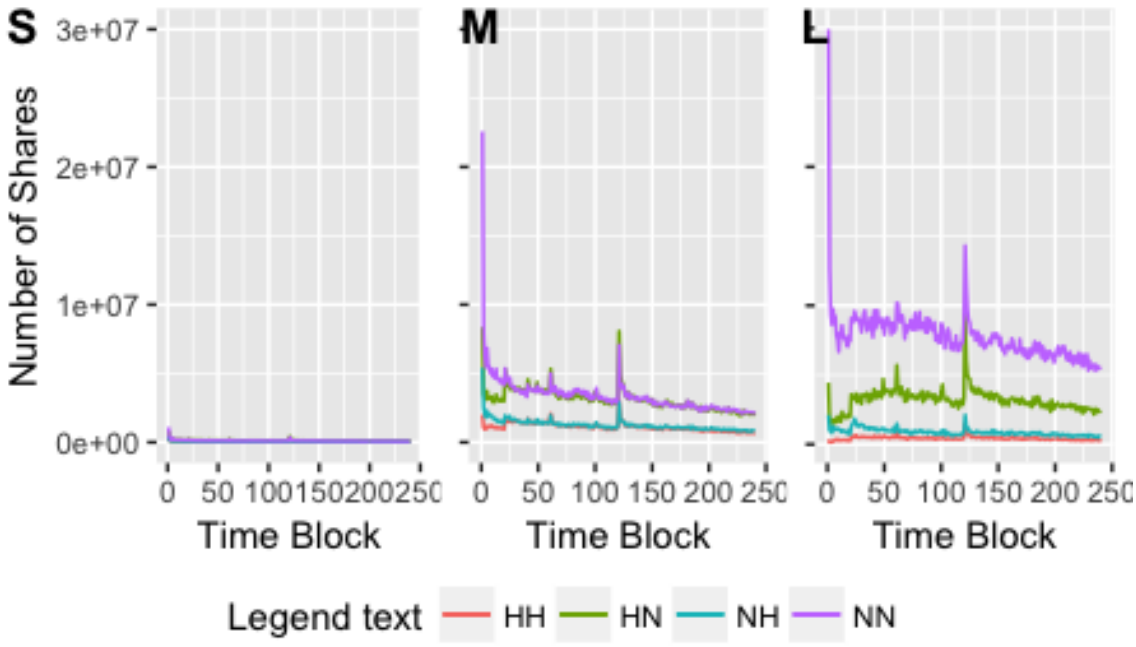

Figure 3.27: *A comprehensive shares graph for time 9:30-10:30*

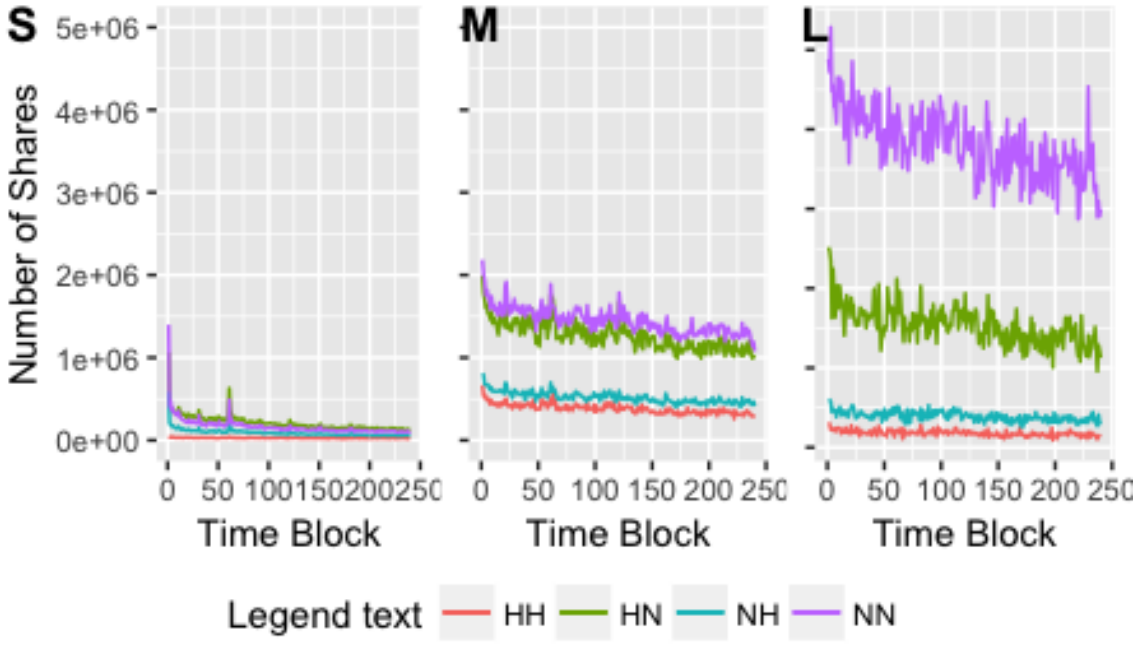

Figure 3.28: *A comprehensive shares graph for time 11:30-12:30*

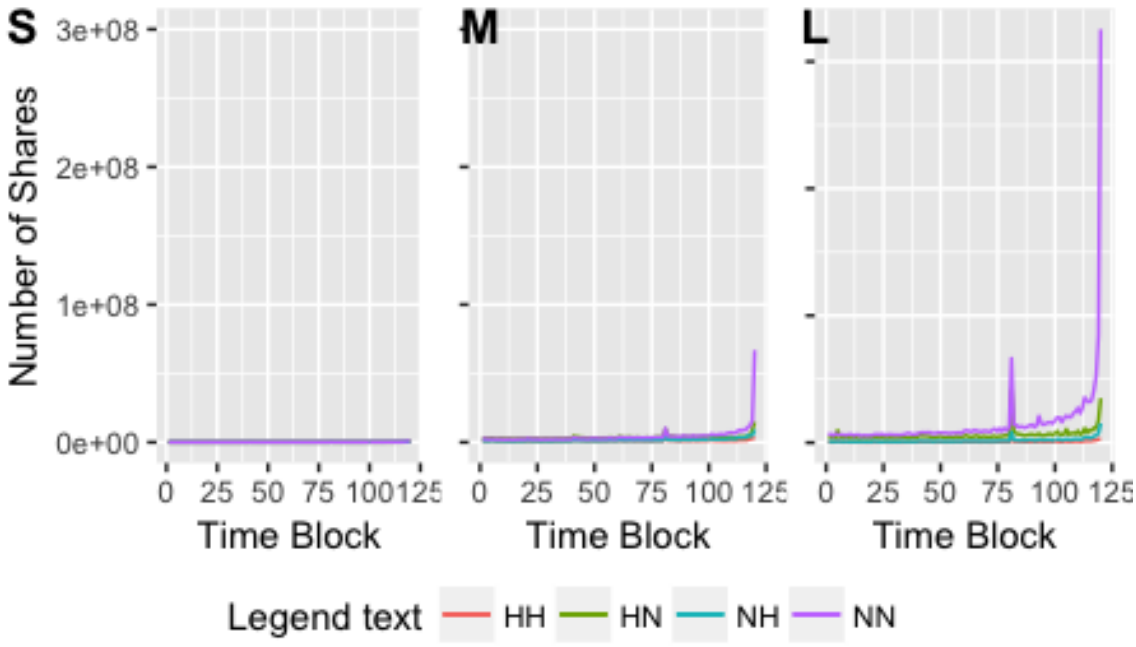

Figure 3.29: *A comprehensive shares graph for time 3:30-4:00*

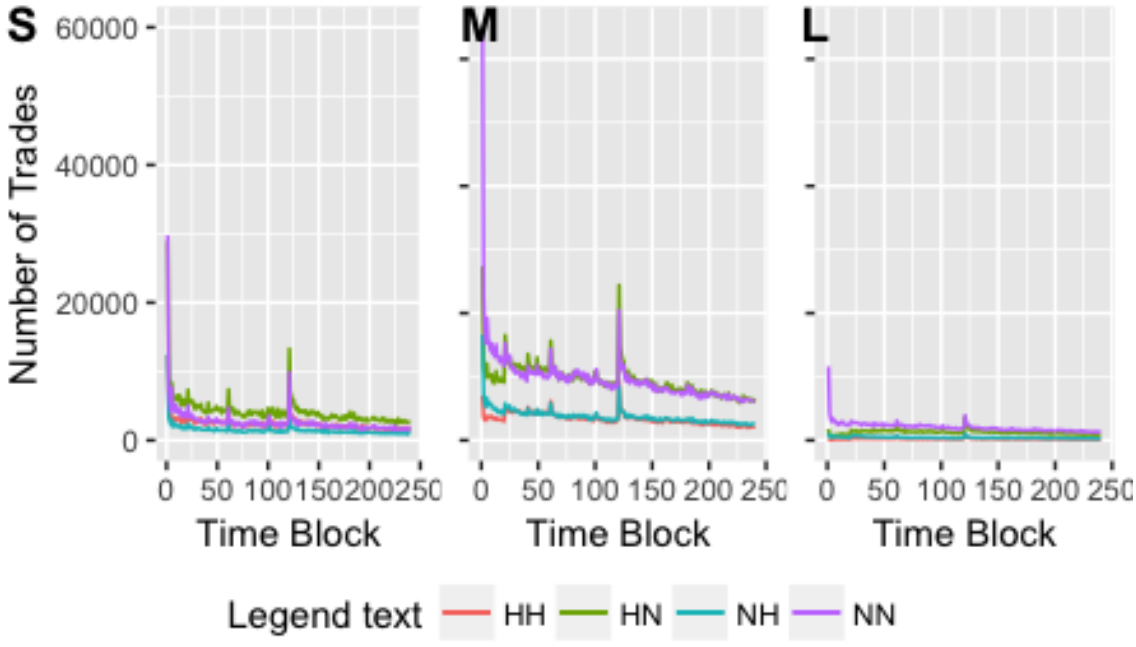

Figure 3.30: *A comprehensive trades graph for time 9:30-10:30*

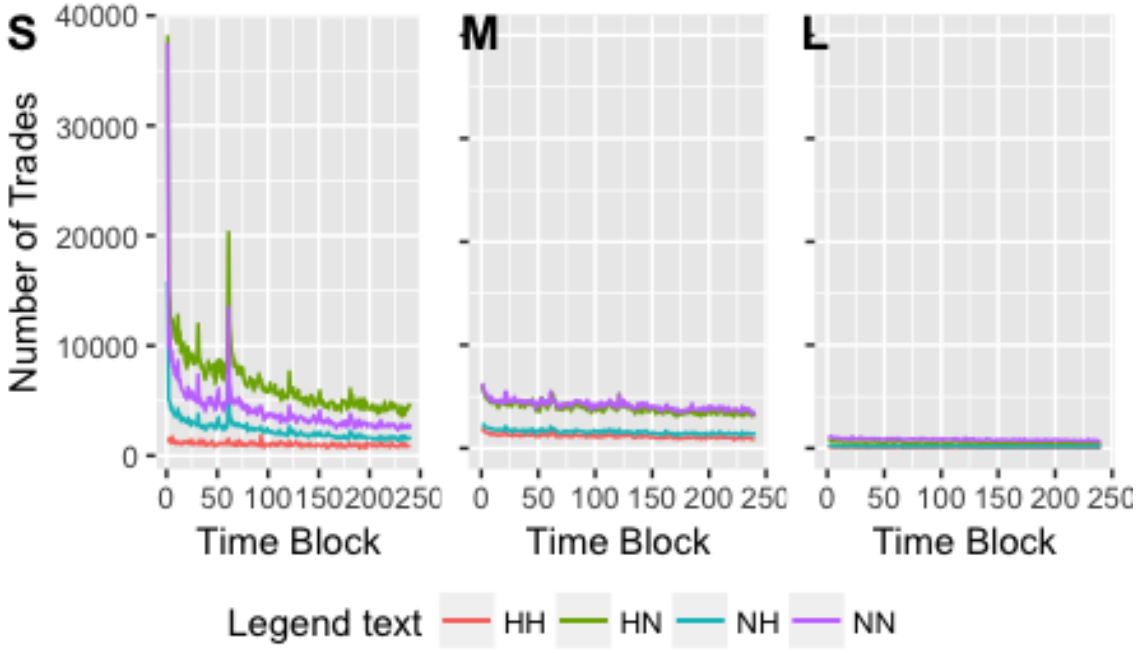

Figure 3.31: *A comprehensive trades graph for time 11:30-12:30*

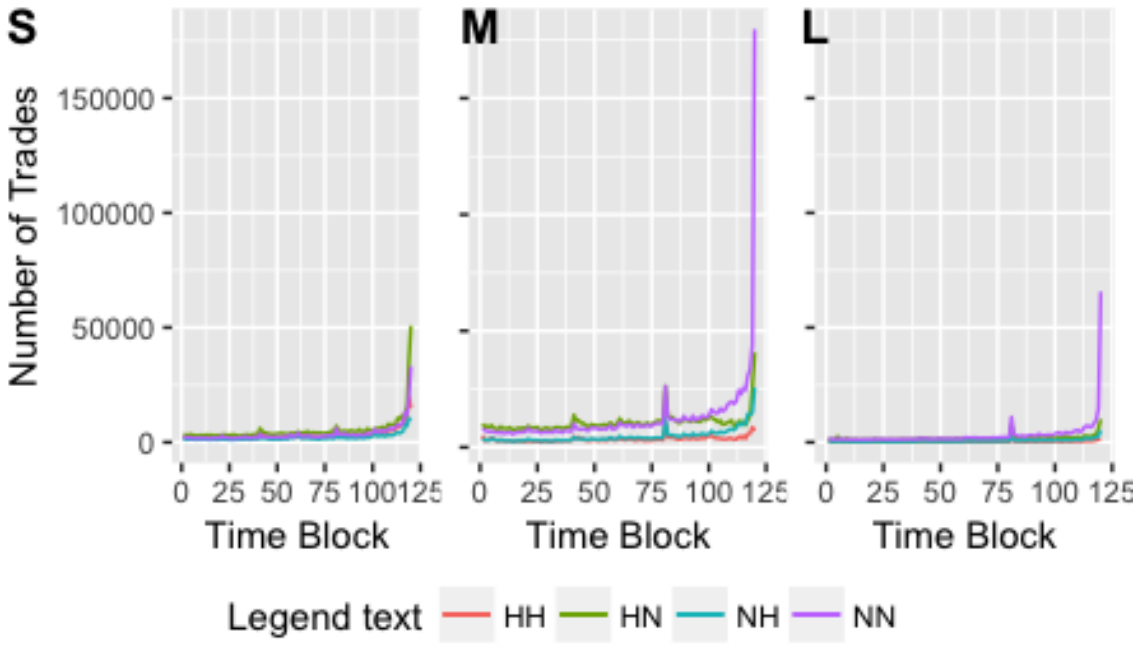

Figure 3.32: *A comprehensive trades graph for time 3:30-4:00*

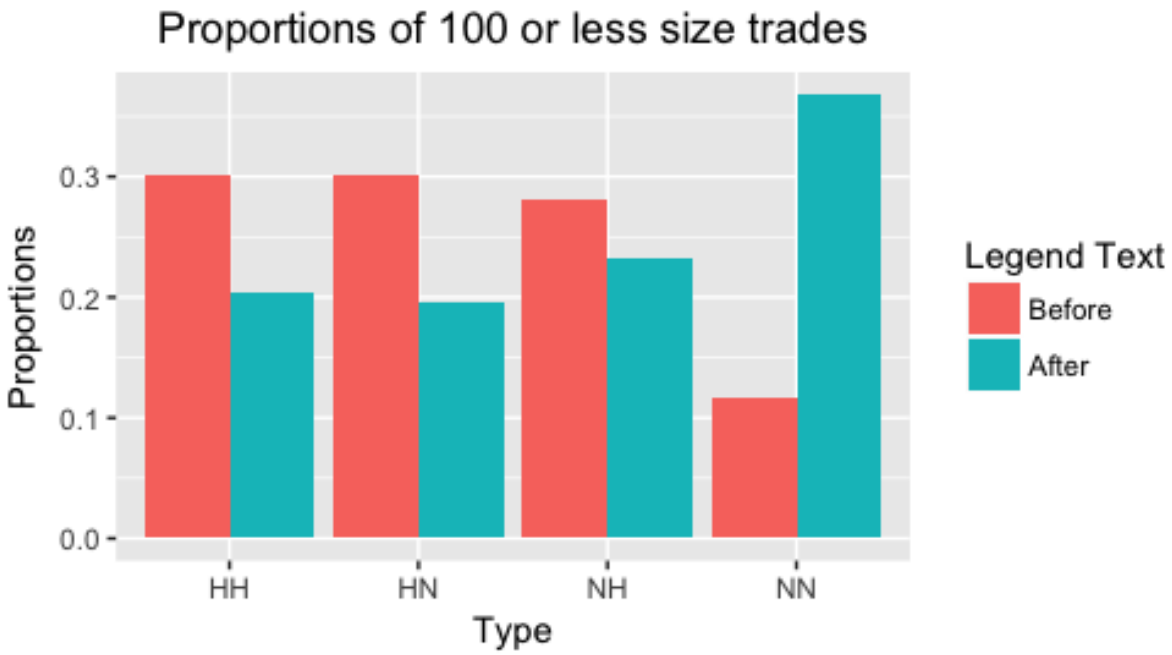

Figure 3.33: *Proportion of 100 or less size trades before and after large trades*

Figure 3.34: *Proportion of small size trades before and after large trades*

	type	price	stock	Milis	BS	NoShares	Date	N
7245134	NH	20.48	INTC	57054985	Sel	5	120409	13749
7245135	NH	20.48	INTC	57054991	Sel	11	120409	13750
7245136	NH	20.48	INTC	57054996	Sel	39	120409	13751
7245137	NH	20.48	INTC	57055007	Sel	10	120409	13752
7245138	NH	20.48	INTC	57055986	Sel	33	120409	13753
7245139	NH	20.48	INTC	57056000	Sel	29	120409	13754
7245140	NH	20.48	INTC	57056010	Sel	11	120409	13755
7245141	NH	20.48	INTC	57056986	Sel	14	120409	13756
7245142	NH	20.48	INTC	57056997	Sel	29	120409	13757
7245143	NH	20.48	INTC	57057008	Sel	6	120409	13758
7245144	NH	20.48	INTC	57057009	Sel	16	120409	13759
7245145	NH	20.48	INTC	57057017	Sel	5	120409	13760
7245146	NH	20.48	INTC	57058028	Sel	9	120409	13761
7245147	NH	20.48	INTC	57058037	Sel	8	120409	13762
7245148	NH	20.48	INTC	57058049	Sel	24	120409	13763
7245149	NH	20.48	INTC	57058063	Sel	10	120409	13764
7245150	NN	20.47	INTC	57058580	Buy	13500	120409	13765

Figure 3.35: *NASDAQ HFT dataset before a trade of 13,500 shares*

	type	price	stock	Milis	BS	NoShares	Date	N
7245275	NN	20.46	INTC	57194058	Buy	900	120409	13890
7245276	NH	20.47	INTC	57194107	Sel	24	120409	13891
7245277	NH	20.47	INTC	57194988	Sel	38	120409	13892
7245278	NH	20.47	INTC	57195106	Sel	14	120409	13893
7245279	NH	20.47	INTC	57195114	Sel	2	120409	13894
7245280	NH	20.47	INTC	57195116	Sel	18	120409	13895
7245281	NH	20.47	INTC	57195988	Sel	15	120409	13896
7245282	NH	20.47	INTC	57196001	Sel	23	120409	13897
7245283	NH	20.47	INTC	57196106	Sel	13	120409	13898
7245284	NH	20.47	INTC	57196115	Sel	16	120409	13899
7245285	NH	20.47	INTC	57196988	Sel	27	120409	13900
7245286	NH	20.47	INTC	57197109	Sel	11	120409	13901
7245287	NH	20.47	INTC	57197114	Sel	25	120409	13902
7245288	NN	20.46	INTC	57198839	Buy	9300	120409	13903
7245289	HN	20.46	INTC	57198866	Buy	20	120409	13904
7245290	NH	20.47	INTC	57205869	Sel	200	120409	13905
7245291	NN	20.47	INTC	57206711	Sel	118800	120409	13906

Figure 3.36: *NASDAQ HFT Dataset before a trade of 118,800 shares*

	type	price	stock	Milis	BS	NoShares	Date	N
7245150	NN	20.47	INTC	57058580	Buy	13500	120409	13765
7245151	HN	20.47	INTC	57058581	Buy	30	120409	13766
7245152	HN	20.47	INTC	57058582	Buy	530	120409	13767
7245153	HN	20.47	INTC	57058583	Buy	870	120409	13768
7245154	HN	20.47	INTC	57058584	Buy	2284	120409	13769
7245155	NH	20.47	INTC	57058585	Sel	100	120409	13770
7245156	HH	20.47	INTC	57058862	Sel	800	120409	13771
7245157	NH	20.47	INTC	57059065	Sel	26	120409	13772
7245158	NH	20.47	INTC	57059080	Sel	11	120409	13773
7245159	NH	20.47	INTC	57059086	Sel	11	120409	13774
7245160	NH	20.47	INTC	57059098	Sel	14	120409	13775
7245161	NH	20.47	INTC	57059386	Sel	265	120409	13776
7245162	HH	20.47	INTC	57059394	Sel	722	120409	13777

Figure 3.37: *NASDAQ HFT dataset after a trade of 13,500 shares*

	type	price	stock	Milis	BS	NoShares	Date	N
7245291	NN	20.47	INTC	57206711	Sel	118800	120409	13906
7245292	NN	20.47	INTC	57206715	Sel	11145	120409	13907
7245293	HH	20.47	INTC	57206716	Sel	1677	120409	13908
7245294	HH	20.47	INTC	57206717	Buy	57	120409	13909
7245295	HH	20.47	INTC	57206719	Buy	23	120409	13910
7245296	NH	20.47	INTC	57206729	Buy	4400	120409	13911
7245297	NH	20.47	INTC	57206733	Buy	100	120409	13912
7245298	NH	20.47	INTC	57206773	Buy	2200	120409	13913
7245299	NH	20.47	INTC	57206809	Buy	100	120409	13914
7245300	NH	20.47	INTC	57206821	Buy	1100	120409	13915
7245301	NH	20.47	INTC	57206830	Buy	268	120409	13916
7245302	HN	20.47	INTC	57206831	Sel	134	120409	13917
7245303	HN	20.47	INTC	57206832	Sel	134	120409	13918
7245304	HN	20.47	INTC	57206833	Sel	134	120409	13919
7245305	HN	20.47	INTC	57206947	Sel	134	120409	13920

Figure 3.38: *NASDAQ HFT dataset after a trade of 118,800 shares*

Chapter 4

Discussion

Usually, when people analyze stock data, they are more likely to focus on price prediction. However, this report focuses on high-frequency trading and the behavior of high-frequency traders. High-frequency trading often uses powerful computers to transact a large number of trades at a high speed. It has very complicated algorithms to execute the trades based on market conditions. High-frequency trading becomes more popular because it can help companies reduce operating costs.

Having the massive stock data, we try to figure out the behavior of high-frequency traders. We make a large number of plots to visualize the data, trying to explore and investigate the trading patterns of high-frequency traders. In conclusion, High peaks are more likely to appear in two time periods: when the stock market just opens and is about to close. Non-high-frequency traders tend to have large number of trades at the beginning time and end time of the stock market trading day. High-frequency traders are more sensible when trading; they are not affected as much by the time opening and closing times. Except for those two time periods, non-high-frequency traders' trading volume is more volatile than high-frequency traders'. Janky trades—those of size less than 100 shares—may be more common done by high-frequency traders. Large trades have large number of shares but most trades are medium-sized.

Since hybridized threshold clustering method has been proposed as a way of clustering large-to-massive datasets, we apply it to our dataset to identify the high-frequency traders' patterns. While this clustering method is feasible for this size of dataset, because of the inherent variability in the covariates, and because of the limited covariate information provided by the data, we were not able to glean much information from this clustering method. Further study will be needed.

For the large trades behavior, high-frequency trades are more likely to perform small trades before large trades. Janky small trades are more common done by high-frequency traders. In some situations, there may be a large number of janky trades performed before a large trade occurs. After the large trade, the janky trades are less common.

Bibliography

David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

Jonathan Brogaard, Terrence Hendershott, and Ryan Riordan. High-frequency trading and price discovery. *The Review of Financial Studies*, 27(8):2267–2306, 2014.

Pasi Fränti and Juha Kivijärvi. Randomised local search algorithm for the clustering problem. *Pattern Analysis & Applications*, 3(4):358–369, 2000.

Pasi Fränti, Juha Kivijärvi, Timo Kaukoranta, and Olli Nevalainen. Genetic algorithms for large-scale clustering problems. *The Computer Journal*, 40(9):547–554, 1997.

Pasi Fränti, Juha Kivijärvi, and Olli Nevalainen. Tabu search algorithm for codebook generation in vector quantization. *Pattern Recognition*, 31(8):1139–1148, 1998.

Joel Hasbrouck and Gideon Saar. Low-latency trading. *Journal of Financial Markets*, 16 (4):646–679, 2013.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*, volume 2. Springer series in statistics New York, NY, USA:, 2009.

Terrence Hendershott, Charles M Jones, and Albert J Menkveld. Does algorithmic trading improve liquidity? *The Journal of Finance*, 66(1):1–33, 2011.

Michael J Higgins, Fredrik Sävje, and Jasjeet S Sekhon. Improving massive experiments with threshold blocking. *PNAS*, 2015.

Mary Inaba, Naoki Katoh, and Hiroshi Imai. Applications of weighted voronoi diagrams and randomization to variance-based k-clustering. In *Proceedings of the tenth annual symposium on Computational geometry*, pages 332–339. ACM, 1994.

Lior Rokach and Oded Maimon. Clustering methods. *Data mining and knowledge discovery handbook*, pages 321–352, 2005.

Michael Steinbach, Levent Ertöz, and Vipin Kumar. The challenges of clustering high dimensional data. *New Directions in Statistical Physics*, pages 273–309, 2004.

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. Data mining cluster analysis: basic concepts and algorithms. *Introduction to data mining*, 2013.