



PhD-FDEF-2022-012
The Faculty of Law, Economics and Finance

DISSERTATION

Defence held on 06/07/2022 in Luxembourg
to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU
LUXEMBOURG

EN SCIENCES FINANCIÈRES

by

Vladimir LEVIN

Born on the 7th of January, 1992 in Kineshma (Russia)

ESSAYS ON

MARKET MICROSTRUCTURE AND FINANCIAL
MARKETS STABILITY

Dissertation defence committee:

Prof. Dr. Roman Kräussl, Dissertation Supervisor
Full Professor, Université du Luxembourg, Luxembourg

Prof. Dr. Ulf von Lilienfeld-Toal, Chairman
Full Professor, Université du Luxembourg, Luxembourg

A-Prof. Dr. Alexander Skupin, Vice Chairman
Associate Professor, Université du Luxembourg, Luxembourg

Prof. Dr. Jérôme Dugast
Professor, Université Paris Dauphine - PSL, France

Prof. Dr. Hans Degryse
Professor, KU Leuven, Belgium

Acknowledgements

This thesis would not have been possible without the support of many people. Many thanks to my supervisors, Roman Kräussl and Jorge Gonçalves, who read my numerous revisions and helped make some sense of the confusion. I have benefited greatly from your wealth of knowledge and meticulous editing. I am extremely grateful that you took me on as a student and continued to have faith in me over the years.

Thank you to my committee members, Ulf von Lilienfeld-Toal, Alexander Skupin, Jérôme Dugast, and Hans Degryse, who offered guidance and support. I also thank my colleagues in the Department of Finance and Laurent Mombaerts from Systems Control Group of the University of Luxembourg for reading the papers that are part of this thesis and providing their valuable comments and feedback.

During my Ph.D., I have attended some great summer schools and had an honor of discussing the research with people like Albert Menkveld, Thierry Foucault, Yacine Aït-Sahalia, Per Mykland, Tarun Chordia, Rajnish Mehra, Christian Brownlees, and others. I am fortunate to have been a part of the #fincap project with my colleague Dr. Gabriel Kaiser. The project resulted in a paper that is currently under the “revise-and-resubmit” status in the Journal of Finance.

And finally, thanks to my wife, parents, and friends who endured this long process with me, always offering support and love.

Contents

General Introduction	6
1 Dark Trading and Financial Markets Stability	9
1.1 Introduction	9
1.2 Data and Descriptive Statistics	12
1.2.1 Data sources	12
1.2.2 Mini-flash crash identification	13
1.2.3 Midpoint Extended Life Order	15
1.2.4 Measures of liquidity and order imbalance	18
1.2.5 Measures of algorithmic trading	19
1.2.6 Summary statistics	20
1.3 Empirical Approach and Results	22
1.3.1 Mini-flash crashes	22
1.3.2 Liquidity provision measures	27
1.3.3 Robustness	28
1.4 Conclusion	29
References	31
Tables and Figures	32
1.A Appendix	43
1.A.1 M-ELO order specifications	43
1.A.2 Stocks and ETFs selected for panel regressions	44
1.A.3 Tests for the instruments	45
2 Dark Pools and Price Discovery in Limit Order Markets	46
2.1 Introduction	46
2.2 Model	49
2.2.1 Limit Order Book and Dark Pool	49
2.2.2 Agents and Strategies	51
2.2.3 Optimal Order Submission	52
2.3 Equilibrium	54

2.3.1	Single-Venue Market Equilibrium	54
2.3.2	Multiple-Venue Market Equilibrium	57
2.4	Venue Competition	60
2.5	Empirical Implications	61
2.5.1	Price Discovery	61
2.5.2	Market Quality	63
2.5.3	Welfare	66
2.6	Conclusion	67
	References	69
	Tables and Figures	70
3	Machine Learning and Market Microstructure Predictability	86
3.1	Introduction	86
3.2	Literature Review	88
3.3	Methodology	91
3.3.1	Linear model	91
3.3.2	Light Gradient Boosted Machine (LGBM)	92
3.3.3	Neural Networks	94
3.3.4	Accuracy metrics for performance comparison	96
3.4	Data	97
3.4.1	Time and dollar-volume bars	98
3.4.2	Features and targets construction	99
3.5	Empirical Results	103
3.5.1	Returns	104
3.5.2	Mini-flash crashes	105
3.5.3	Quoted spread	107
3.5.4	Realized volatility	108
3.5.5	The most important features	109
3.6	Conclusions and further directions	111
	References	116
	Tables and Figures	117
3.A	Appendix	137
3.A.1	Nasdaq order book messages	137
3.A.2	Details on Microstructure Measures employed	138
	General Conclusion	140

List of Tables

1.1	Descriptive statistics of the sample firms	32
1.2	Descriptive statistics of the identified mini-flash crashes	33
1.3	The effect of M-ELO trading on the average weekly number of mini-flash crashes	34
1.4	The effect of M-ELO trading on crash characteristics	35
1.5	The effect of M-ELO trading on market liquidity	36
1.6	Alternative M-ELO activity specifications	37
1.7	The effect of M-ELO trading on the number of mini-flash crashes for big and small stocks and outlier-robust effects	38
A.1	Properties of the instrumental variables	45
2.1	Available order types	70
3.1	Descriptive statistics of the features (2019, time bars)	117
3.2	Descriptive statistics of the features (2020, time bars)	118
3.3	Descriptive statistics of the features (2019, volume bars)	119
3.4	Descriptive statistics of the features (2020, volume bars)	120
3.5	R^2_{oos} for pooled regressions with time bars	121
3.6	R^2_{oos} for pooled regressions with dollar volume bars	122
3.7	Diebold-Mariano test for pooled regressions on one-period returns (time bars)	123
3.8	Diebold-Mariano test for pooled regressions on flash crashes (time bars)	124
3.9	Diebold-Mariano test for pooled regressions on quoted spread (time bars)	125
3.10	Diebold-Mariano test for pooled regressions on realized volatility (time bars)	126
3.11	Diebold-Mariano test for pooled regressions on one-period returns (volume bars)	127
3.12	Diebold-Mariano test for pooled regressions on flash crashes (volume bars)	128
3.13	Diebold-Mariano test for pooled regressions on quoted spread (volume bars)	129
3.14	Diebold-Mariano test for pooled regressions on realized volatility (volume bars)	130
A.1	Nasdaq order book messages	137

List of Figures

1.1	Mini-flash crashes, quoted spread, and book depth throughout the sample	39
1.2	Example of the mini-flash crash	40
1.3	Share of M-ELO executions	41
1.4	Densities of lit (visible) order sizes and M-ELO order sizes	42
2.1	Dark trading in USA and Europe	71
2.2	Limit orders in the LOB	72
2.3	Procedure to obtain an equilibrium in a single-venue market	73
2.4	The extensive form of the trading game in the single-venue market	74
2.5	Procedure to obtain an equilibrium in a multiple-venue market	75
2.6	The extensive form of the trading game in the multiple-venue market	76
2.7	Equilibrium order submission strategies composition	77
2.8	Order Migration for various dark pool availability and information asymmetry .	78
2.9	The probability that the dark order came from an uninformed trader	79
2.10	Price discovery for different levels of α	80
2.11	Average quoted bid-ask spread	81
2.12	Average market depth	82
2.13	Average fill rate	83
2.14	Aggregate welfare	84
2.15	Aggregate welfare of traders with different information sets	85
3.1	Feed-forward Neural Network	131
3.2	LSTM Neural Network	132
3.3	Feature importance for return forecasts	133
3.4	Feature importance for flash crash forecasts	134
3.5	Feature importance for quoted spread forecasts	135
3.6	Feature importance for realized volatility forecasts	136

General introduction

Market microstructure encompasses various fields developed by the researchers from different disciplines like economics and finance, data science, mathematics and physics. The following is an introduction to the topics relevant to this thesis.

Financial markets can be organized in different ways. The two stylized market structures that are particularly relevant for the context of this thesis are an exchange or transparent market and an over-the-counter market or a dark pool. An exchange is a venue that handles trading via the electronic continuous double-sided auction. Various orders (to buy at the market/limit, to sell, to modify the existing order) are collected by the exchange and displayed to all market participants in the form of the limit order book (LOB). The LOB is fully transparent and any participant can observe the available volume at any price at a certain moment in time. The order flow brings changes to the LOB, and the state of the LOB is a crucial piece of information for designing trading strategies.

On the other hand, over-the-counter markets, which are essentially based on bilateral negotiations, vary in the degree of opacity. Dark pools represent one of the most popular structures of over-the-counter markets. They collect buy and sell orders and match them based on volume, but not the price. The price of the resulting transaction is determined from an external source, for example, an exchange.

There is an ongoing debate in the industry and also in the academic world about the positive and negative effects of the growing volume of dark trading. In the first and the second chapters, I look into the implications of dark trading for market stability, liquidity, and welfare. The first chapter is an empirical work that employs the quasi-natural experiment in order to understand the influence of dark trading on the overall market quality and stability. The second chapter presents a theoretical model of the limit order book and an attached dark pool where I observe the changes in the market parameters as the availability of the dark pool changes.

Another important aspect of the current markets is the speed arms race. In recent years, with digitalization and the evolution of electronic commerce, it has become common for financial market participants to process their transactions fully electronically. The high speed at which the transactions could be processed and settled gave name to the new trading style known as high-frequency trading (HFT). High-frequency traders (HFTs) usually try to locate as close to the exchange matching engine as possible, as they value speed advantages up to nanoseconds. They buy and sell quickly and do not hold a position for a long time.

In the current markets, the volume share traded by HFTs certainly makes them the key players in equity markets. However, the activity of HFTs that can result in higher volatility of intraday prices, variation of trading volumes, and evaporation of the liquidity in periods of stress, has recently been under the radar of regulating authorities.

The HFT aspect of the modern market microstructure is addressed in the first and the third chapters of the current thesis. The first chapter assesses the degree to which the speed bump effect in a particular hidden order on Nasdaq influences market stability, volatility, and liquidity. The third chapter describes the use of machine learning techniques in order to predict short-term market stability, volatility and liquidity. The proxy for the HFT is one of the important microstructure features used in models training.

The present doctoral thesis consists of three main chapters. The chapters of the thesis can be considered independently. Each of the three chapters raises a research question, reviews the related literature, proposes a method for the analysis, and, finally, reports results and conclusions.

Chapter 1 is entitled *Dark Trading and Financial Markets Stability* and it is based on a working paper co-authored with Prof. Dr. Jorge Gonçalves and Prof. Dr. Roman Kräussl. This paper examines how the implementation of a new dark order – Midpoint Extended Life Order (M-ELO) on Nasdaq – impacts financial markets stability in terms of occurrences of mini-flash crashes in individual securities. We use high-frequency order book data and apply panel regression analysis to estimate the effect of dark order trading activity on market stability and liquidity provision. The results suggest a predominance of a speed bump effect of M-ELO rather than a darkness effect. We find that the introduction of M-ELO increases market stability by reducing the average number of mini-flash crashes, but its impact on market quality is mixed.

Chapter 2 is entitled *Dark Pools and Price Discovery in Limit Order Markets* and it is

a single-authored work. This paper examines how the introduction of a dark pool impacts price discovery, market quality, and aggregate welfare of traders. I use a four-period model where rational and risk-neutral agents choose the order type and the venue and obtain the equilibrium numerically. The comparative statics on the order submission probability suggests a U-shaped order migration to the dark pool. The overall effect of dark trading on market quality and aggregate welfare was found to be positive but limited in size and depended on market conditions. I find mixed results for the process of price discovery. Depending on the immediacy need of traders, price discovery may change due to the presence of the dark venue.

Chapter 3 is entitled *Machine Learning and Market Microstructure Predictability* and it is another single-authored piece of work. This paper illustrates the application of machine learning to market microstructure research. I outline the most insightful microstructure measures, that possess the highest predictive power and are useful for the out-of-sample predictions of such features of the market as liquidity volatility and general market stability. By comparing the models' performance during the normal time versus the crisis time, I come to the conclusion that financial markets remain efficient during both periods. Additionally, I find that high-frequency traders activity is not able to forecast accurately neither of the market features.

Chapter 1

Dark Trading and Financial Markets Stability

1.1 Introduction

For a couple of decades, market participants have been spending massive resources to obtain quick access to the richest data and invested in technologies to execute trades as fast as possible. According to Easley, López de Prado, and O'Hara (2012), for the period from 2009 to 2012, the share of high-frequency trading (HFT) firms have risen to more than 70% in the U.S. equity markets and approached 50% of the volume in futures markets. Since then, the percentage of HFT firms had increased further. Generally, HFTs use computer algorithms to look at patterns of prices, volumes, and past trading activity and react to any changes in those patterns in a matter of micro- or even nanoseconds. Some of them would not consider themselves as investing in fundamental information, but rather acquiring information about market dynamics and liquidity. Because fast actions of one algorithm may trigger responses of many others, small mispricing can rapidly self-reinforce itself and occasionally cause flash crashes in securities' prices.

Golub et al. (2012) study the increase in the number of mini-flash crashes in individual securities between 2006 and 2011 and suggest that HFT causes those crashes. Leal et al. (2016) build an agent-based model to study how the interplay between low- and high-frequency trading affects asset price dynamics. They find that the presence of HFT increases market volatility and plays a fundamental role in the generation of flash crashes. On the other hand, Kirilenko

et al. (2017) examine the structure of the E-mini S&P 500 stock index futures market on May 6, 2010, and observe that trading patterns of HFT did not change when prices fell during the Flash Crash.

Biais and Foucault (2014) and Biais et al. (2015) propose to create a segment of slow-friendly markets but to leave room for investment in fast trading technology. In the spirit of this recommendation, exchanges started to introduce technology-based solutions to protect the interests of long-term investors. Those solutions were implemented in the form of latency delays and are commonly known as “speed bumps”. The Investors Exchange (IEX) applied the first such measure by introducing a 350-microsecond delay to all incoming and outgoing correspondences. Hu (2018) observes improvements in market functioning around the period when IEX became a national securities exchange. Moreover, he documents a positive impact of such speed bumps on market quality in terms of tighter spreads and improved liquidity. Several other exchanges followed the example of IEX and applied for the introduction of such a delay to the Securities and Exchange Commission (SEC).

Dark trading is an alternative approach to slow down markets. In addition, dark orders also hide trading intentions. When trading large amounts of stock using visible market or limit orders, one cannot prevent the price to be moved. To reduce price impact, a trader can submit the order to the dark pool and often even receive a better execution price. The downside of going to the dark pool is execution uncertainty since there is no guarantee that a trader will find a counterparty.

Major exchanges nowadays run their own dark pools where the execution price is referenced by the current mid-price (the average of the best bid and the best ask prices). It is difficult, however, to hide dark orders from HFT firms. They use their speed advantage to submit hidden orders inside the spread and quickly cancel them if they do not execute right away. If HFTs identified hidden orders, they could easily manipulate the best quotes to transact with those orders at comfortable prices.

In 2018, the Nasdaq exchange came up with a solution to improve dark orders and shield them from HFTs. Introduced on March 12, 2018, the Midpoint Extended Life Order (M-ELO) is targeted toward long-term investors. Anonymity and confidentiality of M-ELO are the critical tools to prevent potentially predatory counterparties from determining intentions and using that information to generate short-term profits at the expense of slow traders. This

order becomes executable 500 milliseconds after submission and does not interact with other Nasdaq dark orders that have not met the 500 milliseconds holding period requirement. On May 11, 2020, Nasdaq revisited the design of the M-ELO order and decided to reduce the holding period to 10 milliseconds. The exchange motivated this change by optimization in M-ELO opportunities and execution. The decrease in the holding period was expected to open up M-ELO to use cases that were previously unavailable¹.

Despite increasing volumes traded in dark pools, there is still limited theoretical work discussing the effect of hidden orders on market quality, stability, and price discovery. Boulatov and George (2013) build a model where the strategies of informed traders can be adjusted in response to the visible and hidden liquidity on the market. They analyze venue and order choices of traders and find that hidden liquidity has a beneficial impact on market quality due to increased competition among informed market participants.

In contrast to Boulatov and George (2013), Zhu (2014) argues that adding a dark pool alongside the exchange decreases its liquidity. His model suggests that, since informed orders are much more correlated than uninformed ones, informed traders would rather choose the lit venue to avoid low execution probabilities in the dark pool. This relatively high presence of informed trading on the lit exchange naturally increases price discovery which, in turn, leads to reduced liquidity. Buti et al. (2017) numerically solve a discrete model in which traders decide to submit their order to either an exchange or a dark pool. The authors obtain a set of equilibrium order submission probabilities and show that the introduction of a dark pool alongside the exchange widens bid-ask spread and reduces the depth available around the midquote. Those negative changes in market quality are partially mitigated when the initial liquidity of the limit order book increases.

Overall, theoretical works propose mixed results about the effects of dark trading on market quality and price discovery. It is still a challenge for regulators to decide on the degree of control of dark trading. Moreover, the aspect of market stability received much less attention. However, it remains a relevant topic in the current times of algorithmic trading proliferation since fragile markets may undermine investors' trust in the financial market system.

This paper adds to the existing literature on dark trading, speed bumps and also expands

¹More information on the rule change may be found at <https://www.nasdaq.com/articles/the-midpoint-extended-life-order-m-elo%3A-m-elo-holding-period-2020-02-13>

on its relation to market stability. It aims to identify the degree to which M-ELO is used in securities trading and its impact on the number of mini-flash crashes during a continuous trading period. We also investigate the association between M-ELO trading and market liquidity. The analysis employs high-frequency trades and quotes data from Nasdaq to identify mini-flash crashes in individual securities and to relate crash occurrences to the intensity of dark M-ELO trading through panel regression analysis. Recently implemented change in the order design allows also to disentangle the darkness and the speed bump effects of the M-ELO. Our results suggest a strong relationship between M-ELO volumes and measures of market quality and stability. This relationship is more pronounced through the speed bump effect of M-ELO rather than through the dark order effect. Overall, M-ELO is associated with fewer flash crashes, greater visible volumes in the limit order book, but widened spreads.

The remainder of the paper is organized as follows. Section 1.2 describes the data sources of intra-day trading and M-ELO volumes. It also presents measures of liquidity and summary statistics. Section 1.3 presents the methodology, empirical results, and describes various robustness checks. Section 1.4 concludes.

1.2 Data and Descriptive Statistics

1.2.1 Data sources

The Order Book Message data come from Nasdaq’s historical ITCH. This data set contains time-stamped in nanoseconds order submissions, executions, cancellations, and modifications on the Nasdaq equity market. The data, however, do not identify market participants and their activity in the dark. Submissions of hidden orders of any kind are not reported, while the executions are visible for all order types.

The data allow us to directly observe liquidity provision on each depth level of the limit order book at any time. The sample period covers almost 3 years of trading from January 2, 2018, to December 31, 2020. We consider only large-cap constituents of the set of firms traded on Nasdaq, where the trading activity is considerably high. We, therefore, in line with similar analyses of Andersen et al. (2001) and Brogaard et al. (2018), preserve the sufficiently large number of observations by focusing our analysis on the largest firms.

The limitations of this data are straightforward and similar to those that previous research

encountered (see, e.g., Carrion (2013); Brogaard et al. (2014); O’Hara et al. (2014); Brogaard et al. (2018)). We do not observe individual HFT activity as well as trading activity on other venues. Trades on Nasdaq account for, on average, 33% of trading activity for Nasdaq listed stocks, about 12.5% for NYSE stocks, and 16% for ARCA stocks. Despite the high fragmentation nature of financial markets, we share the reasoning of Brogaard et al. (2018) that liquidity transfers to other venues are unlikely due to the short period of interest and overall similar liquidity provision rules among exchanges. Thus, we argue that although the results obtained could not readily expand to other exchanges, still they should be taken into account in the matters of market design.

The M-ELO order became available on March 12, 2018, and quickly gained its share of trading volume which averages to be around 1.82% of Nasdaq’s total matched volume for the period from 2018 to 2020. Weakly volumes of M-ELO trades come from Nasdaq’s M-ELO Transparency Statistics². This is one of the biggest limitations of the present analysis. By shifting to weekly observations one might lose statistical power in identifying the effect of M-ELO trading. We resort to the fact, however, that the size of the sample is sufficient and the panel structure of the data allows us to obtain robust estimates. Securities’ characteristics to serve as covariates in our analysis were computed from trades data. As an instrument for the potentially endogenous M-ELO trading, we use the daily average price computed from trades data and the security’s price rank obtained from the SEC’s MIDAS Market Structure Metrics database.

1.2.2 Mini-flash crash identification

We identify mini-flash crashes using an approach similar to Brogaard et al. (2018) for extreme price movements (EPM) identification. In their methodology, the trading day is split into 10-second intervals between 9:35 a.m. and 3:55 p.m. The intervals are then ranked by the midquote return magnitude and those with returns exceeding the 99.99th percentile are identified as EPMs.

However, our methodology differs from Brogaard et al. (2018) in several aspects. Firstly, we account for cross-sectional heterogeneity in trading activities for different firms. This is done by switching from a static time interval of 10 seconds to an interval between 0 and 300 seconds.

² Available at <https://www.nasdaqtrader.com>

Only intervals between 10:00 a.m. and 3:30 p.m. are considered, to abstain from the periods of excessive volatility in the markets. The length of the interval depends directly on the number of trades in the security for a day. It is chosen in a way that each time interval contains on average 50 trades. To make sure the turnover is considerably high, we keep only those securities for which the number of trades for the period from 10:00 a.m. to 3:30 p.m. is not less than 3,300. This way, we require each time interval for mini-flash crash identification to contain at least 50 trades on average. As the sample period is quite long, it is challenging to obtain a set of stocks of an appropriate size that satisfy this condition for the straight three years. That is why, for this analysis, we require stocks to satisfy for at least one-third of the sample period. Throughout the period from 2018 to 2020, we identified 196 firms, that satisfy the minimum daily number of trades condition for at least 50 weeks throughout the 3-year period.

Secondly, as flash crashes are known for their subsequent reversals a simple calculation of midquote returns has a flaw of missing those intervals within which the price suddenly jumps and quickly retraces back. To overcome this problem, for each time interval we compute the maximum possible midquote return. In such a way we obtain a series of intra-day extreme returns for a security.

Lastly, for each firm, we identify the intervals containing mini-flash crashes as those where the Z -score for the midquote extreme returns at day t exceeds the value 7. These are the returns that satisfy $r_{t,i} > \mu_t + 7\sigma_t$, where $r_{t,i}$ is the i^{th} extreme return on day t , μ_t is the average extreme return on day t , and σ_t is the standard deviation of extreme returns on day t for this particular security.

The procedure gets a total of 10,113 mini-flash crashes with 56.94% of them being negative. This identification technique is in line with previous works of Golub et al. (2012) and Johnson et al. (2013) that do not make a sharp distinction between crashes and spikes and require the price to move fast and severely. Similar to the approach of Bellia et al. (2020), we study only those mini-flash crashes which possess transitory dynamics. The average price reversal in the next 10 minutes following the end of the crash is 78.1%.

The first panel of Figure 1.1 shows the distribution of mini-flash crashes throughout the sample period. There is no much visual evidence of M-ELO introduction having a significant effect on the frequency of mini-flash crashes. We associate this fact with relatively low trading volumes of M-ELO orders. However, as we show in this work, there is a statistically significant

association between the degree of M-ELO trading and the expected number of mini-flash crashes observed during a week.

The last two panels of Figure 1.1 show the time trends for the median values of the quoted spread and the depth available 10 basis points around the midquote relative to the daily trading volume. The dynamics of the depth suggests market quality is improving during times of active M-ELO trading in the first half of 2018. In 2019, the relative depth is at its highest levels, sometimes exceeding 2% of the daily trading volume, while the spread stays moderate at around 3 basis points. In early 2020 we can clearly see the impact of the COVID-19 crisis. Spreads more than double while fewer shares are available close to the midquote.

Figure 1.1 around here

An example of a mini-flash crash, identified by our approach is presented in Figure 1.2. The crash occurred in the price of Procter & Gamble (PG) on March 21, 2018. Panel A spans the trading during the opening auction, continuous trading period, and closing auction. This example illustrates the typical dynamics of midprice during a mini-flash crash. At 2 p.m., the price experienced a rapid, massive spike of about 1%. In the next five minutes, however, the price dropped more than 1.4% and eventually returned to the region of previous daily consolidation. The price became more volatile during the rest of the day.

Figure 1.2 around here

Panel B shows a zoomed representation of the crash event, where each dot represents a single trade. We note that the crash did not trigger the circuit breaker, even though the return associated with the initial spike had a Z -score of 7. The duration of this mini-flash crash in the price of PG was 26.2 seconds (the initial spike), with a cumulative return of 0.98%, and a trading volume that exceeded \$633 thousand.

1.2.3 Midpoint Extended Life Order

Recent research on dark trading looks at the economics of liquidity provision in the dark. Academics start to distinguish between two types of dark trading. The first, “one-sided” type, reflects dark trading at a single price which, in most cases, is the midquote. It is usually not regarded as a beneficial type of dark trading due to its low execution probability, absence of

profitable market strategies of earning the spread, and limited abilities to hide trading intentions because of susceptibility to probing orders. The other, “two-sided” dark trading, allows dark buy and sell limit orders to exist simultaneously. This type of dark trading is believed to be benign to price discovery and market quality.

Foley and Putninš (2016) and Comerton-Forde et al. (2018) analyze the effect of both types of dark trading by exploiting natural experiments in Canadian and Australian markets. They find that two-sided dark trading reduces quoted, effective, and realized spreads as well as market illiquidity measured by the price impact. On the other hand, they find no evidence that one-sided dark trading affects markets. The dark order studied in our paper can be mainly referred to as one-sided dark trading. However, it possesses some features of the two-sided type as well.

On March 12, 2018, Nasdaq launched a new order type: Midpoint Extended Life Order (M-ELO), which is designed to attract long-term investors to trade with each other at the midpoint of the National Best Bid and Offer (NBBO). M-ELO is a hidden order which interacts only with other M-ELO-type orders³. Because of this, M-ELO orders stay out of the way of the book clearing orders, with the aim to reduce information leakage and to provide a better execution price.

There is a 500 millisecond period called the “Holding Period” before an M-ELO order can be executed. This restriction protects market participants from the negative price impact as well as from adverse selection. If a bid or an ask price moves, M-ELO orders are automatically tagged to the new midquotes but the 500 milliseconds timer does not restart. If that was not the case, then one would expect to see fewer M-ELO orders executed during volatile markets, since prices will move a lot and sometimes exhibit mini-crashes. As the waiting locked-in times cannot be easily extended by the NBBO moves, the documented negative association between the number of matched M-ELOs and the number of mini-flash crashes may not be simply explained by the built-in “protective” features of M-ELOs.

Starting from May 11, 2020, the Nasdaq exchange reduced the holding period of M-ELO orders from 500 to 10 milliseconds. Following the satisfactory M-ELO performance, the exchange decided to increase the opportunity set for its clients. With a 98% reduction of the waiting period, more trading strategies will be able to incorporate the benefits of the M-ELO. For the current research, this rule change is crucial as it helps to disentangle the dark trading effects

³Specifications and more details of M-ELO order type can be found in the Appendix [1.A.1](#).

and the speed bump effects of M-ELO orders.

Figure 1.3 shows that median M-ELO trading reaches 1% of the total matched by Nasdaq volume in the middle of May 2018. In the second half of 2018, it starts to decline which can be associated with the change in the submission fees for this order type. M-ELO was fees-free until May 2018, with the possibility of a one-month extension, given that some trading activity-based milestones were reached. Since June 2018, M-ELO order submission incurs a fee for all stocks. This might be the reason for the gradual decline in relative M-ELO volumes already in the early autumn of 2018.

In 2019, relative M-ELO trading reached its maximum levels. On average M-ELO executions constituted 3.2% of the total amount matched by Nasdaq. The share of M-ELO executions reduced significantly during the first months of 2020. We observe a spike in relative M-ELO volumes right at the time of the design change. As the holding period became less restrictive, more market participants opted for this order. In our analysis, we control for this change by introducing a dummy variable that equals one if the current holding period is 10 milliseconds and zero otherwise. Also, we include time effects into the regression to rule out end of the year effects as well as the COVID-19 related financial crisis.

Figure 1.3 around here

Another remarkable feature of M-ELO is that the sizes of executed orders are usually bigger than the sizes of visible limit orders, submitted to the book. Figure 1.4 plots density curves of sizes of visible limit orders and M-ELO orders. It can be seen that the distribution of sizes for M-ELO dominates the distribution of sizes of visible orders. The fact that M-ELO is associated with bigger order sizes can be a potential channel through which M-ELO activity may impact market stability. HFT would benefit from taking the opposite side relative to market participants who submit large orders. This contrarian trading by HFT may create unreasonable price pressure that can lead to crashes. It is possible that as more market participants opt for non-displayed M-ELO orders, fewer flash crashes occur in the market. The validity of this reasoning is, however, not investigated in this paper, so we remain cautious and do not claim the causality.

Figure 1.4 around here

1.2.4 Measures of liquidity and order imbalance

Liquidity is generally understood as the ability to quickly trade considerable volumes at a low cost. It is a multi-dimensional concept that includes trading costs, depth available to customers placing large orders, speed of execution, and protection against execution risks (Foucault et al. (2013a)).

Our first empirical measure of liquidity is the quoted half spread. It represents a scaled by the midpoint price difference between the lowest ask price (a_t) and the highest bid price (b_t) available at the moment:

$$QS_t = \frac{a_t - b_t}{2m_t} = \frac{a_t - b_t}{a_t + b_t}, \quad (1.1)$$

where $m_t = (a_t + b_t)/2$ is the midprice. The quoted spread for stock-day is time-weighted and based on the local limit order book.

The other measures are the 5-minute realized spread and price impact. They are calculated per trade and then averaged over the trading day. The 5-minute proportional realized spread for the t^{th} transaction is defined as

$$RS_t = d_t \frac{p_t - m_{t+5min}}{m_t}, \quad (1.2)$$

where p_t is the trade price, d_t is the buy-sell indicator that equals +1 if the trade is a buy and -1 if the trade is a sell, and m_{t+5min} is a quote midpoint 5 minutes after the t^{th} trade.

A price impact measure is based on the extent to which a trade generates an adverse reaction in the market price. The midprice tends to rise when buy orders arrive, to an extent that is positively correlated with their size. Symmetrically, it tends to fall in the wake of sell orders. The 5-minute price impact of a t^{th} trade is defined as follows:

$$PI_t = d_t \frac{m_{t+5min} - m_t}{m_t}. \quad (1.3)$$

Our data set allows us to construct various depth measures and incorporate the limit orders beyond the best price levels. Similar to the work of Degryse et al. (2015), we measure the aggregate monetary value of shares offered within a fixed interval around the midpoint. We keep the original notation and refer to this measure as $Depth(X)$. Denote the price level

$j = \{1, 2, \dots, J\}$ on the pricing grid and the midpoint of the local limit order book as m , then

$$\text{Depth Ask}(X) = \sum_{j=1}^J p_j^a \cdot q_j^a \cdot \mathbb{I}\{p_j^a < m(1 + X)\}, \quad (1.4)$$

$$\text{Depth Bid}(X) = \sum_{j=1}^J p_j^b \cdot q_j^b \cdot \mathbb{I}\{p_j^b > m(1 - X)\}, \quad (1.5)$$

$$\text{Depth}(X) = \text{Depth Bid}(X) + \text{Depth Ask}(X), \quad (1.6)$$

where p_j^a (p_j^b) is the price of the limit sell (buy) order at price level j and q_j^a (q_j^b) is the number of shares available at this level. We use the indicator function $\mathbb{I}\{\}$ to determine if a limit order of a certain price is within the required interval around the midquote. The depth measure is expressed in U.S. dollars and calculated for $X = 10$ basis points.

We also expand on measures of order imbalance. Each transaction is labeled as buyer- or seller-initiated based on the information of corresponding limit order submissions. Following Chordia et al. (2002), we compute for each stock the OIBNUM_{*t*} measure of order imbalance, where OIBNUM_{*t*} is the number of seller-initiated trades less the number of buyer-initiated trades on the day t , scaled by the total number of trades.

Further, we calculate an imbalance measure similar to that of Belter (2007). This measure allows us to compare the liquidity supplied to different sides of the book beyond the best price levels. Having the interval of X basis points around the midquote and the price levels $j = \{1, 2, \dots, J\}$, the depth imbalance is defined as follows:

$$DI(X) = \frac{\sum_{j=1}^J \left(\frac{q_j^a}{(p_j^a - m)} \cdot \mathbb{I}\{p_j^a < m(1 + X)\} - \frac{q_j^b}{(m - p_j^b)} \cdot \mathbb{I}\{p_j^b > m(1 - X)\} \right)}{\sum_{j=1}^J (q_j^a + q_j^b) \cdot \mathbb{I}\{1 - X < p_j/m < 1 + X\}}. \quad (1.7)$$

For each stock-day, we compute the average depth imbalance for $X = 10$ basis points. This measure is scaled by the total number of shares available in the given interval X .

1.2.5 Measures of algorithmic trading

It is hard to discriminate between orders placed by humans and orders placed by computer algorithms. A methodology to identify HFT depends heavily on the availability of data. If the data include information on HFT firms, it can be used directly to account for algorithmic

trading (AT) activity. Most financial markets, however, do not provide information on whether an order comes from a human or an algorithm.

In case that HFT cannot be classified exactly, researchers use various proxies to quantify levels of HFT. Those proxies are constructed from trade and order submission data. We use the empirical measure developed in Hasbrouck and Saar (2013) as our major proxy for AT. This measure calculates the intensity of “strategic runs”, which are series of linked messages. The linking results from HFT dynamically submitting and canceling orders to incorporate the latest information into prices.

Following their methodology, we connect a newly submitted limit order to a previously deleted order if the time between the two events does not exceed 100 milliseconds. The newly submitted order should have the same direction and size in shares as the previously deleted one. Only sufficiently long runs of 10 and more linked orders are kept. We scale the sum of durations of all runs, which are allowed to overlap, by the duration of the trading day without the first and the last 30 minutes of trading. Our proxy for the AT activity on day t is defined as follows:

$$AT_t = \frac{1}{5.5 \cdot 3,600} \sum_{j=1}^N T_{jt}, \quad (1.8)$$

where $5.5 \cdot 3,600$ is the total time in seconds from 10:00 a.m. to 3:30 p.m., N is the number of strategic runs on day t , and T_{jt} is the duration in seconds of run j on day t .

1.2.6 Summary statistics

Table 1.1 presents descriptive statistics for the sample stocks. There is considerable variation in average daily price and daily dollar volume traded. The average coefficient of variation of stock extreme returns (σ_r/μ_r) is centered around 0.63 with a relatively small standard deviation. This suggests the firms fall in pretty much the same volatility cohort. The Nasdaq’s share shows what fraction of the consolidated volume in a particular security was matched by the Nasdaq exchange. The mean and the median share of Nasdaq across all listing venues are around 21%.

Table 1.1 around here

Table 1.1 also reports descriptive statistics on liquidity measures and measures of AT. The average daily quoted half spread throughout the sample period is about 4.6 basis points. Five-

minute realized spread and price impact are centered around zero with the first (the third) quartile being about minus (plus) one basis point.

There is a large variation in the dollar value of shares available 10 basis points around the midquote. Its mean of \$1.55 million is higher than the value of its 3rd quartile. The majority of depth available is sell-side volume, as the positive mean value of depth imbalance suggests. The trade-based measure of buy-sell imbalance, however, indicates that there is more buying activity compared to the selling activity. These two observations do not contradict each other since high-level buying activity stimulates liquidity providers to post sell limit orders due to the reduced risk of non-execution.

The proxy of AT based on strategic runs suggests that somebody is engaging in dynamic order submission about 1.8% of the time during the period from 10:00 a.m. to 3:30 p.m. The message-to-trade ratio clearly illustrates the fact that quoting activity nowadays is superior to trading activity. Its mean shows that there are about 40 times more quote update messages than actual trades.

Table 1.2 shows summary statistics for 10,113 identified mini-flash crashes. The Z -score of the return is the value $\frac{r_i - \mu_r}{\sigma_r}$, where r_i is the return during the mini-flash crash, μ_r and σ_r are the mean and standard deviation of maximum interval returns during the day. The threshold for the Z -score was chosen to be equal to 7, to identify extreme price movements. The table reports the average Z -score for the extreme returns around 8.5.

The average duration of a crash of approximately 48 seconds. In fact, 99% of identified mini-flash crashes do not last longer than 164 seconds. The mini-flash crash returns are distributed around -0.08 basis points, which is explained by the fact that only 56.94% of crashes are negative. The median absolute mini-flash crash return is about 58.3 basis points.

Table 1.2 around here

As Table 1.2 reports, mini-flash crashes are remarkable in subsequent price reversals. The 10-minute price reversal after the crash is on average 78%. The number of trades during the crash period is substantially higher than in normal times. With the mean number of trades of 305 and the mean crash duration, one obtains that the average number of trades per second to be 6.4 trades. This is a more than 7 times higher trading intensity compared to the average across the whole sample. The number of shares traded and the dollar volume during a crash tell a

similar story. Also, the size of the trade increases slightly during crash times. The average size of the trade during normal times is around 124 shares, while it increases to 152.1 at the periods where we identify mini-flash crashes.

Table 1.2 also reports summary statistics on the crash aftermath volatility. This is the standard deviation of the extreme returns for the next 30 minutes following the end of the mini-flash crash. It is scaled by the standard deviation of the extreme returns throughout the day. The average volatility after the crash is approximately 16% higher than the volatility for that particular day.

Finally, descriptive statistics on the relative amount of M-ELO trading are provided. The average share of all M-ELO orders relative to all matched by Nasdaq orders is 1.98% for the stocks with identified mini-flash crashes. Also, the size of M-ELO orders is usually larger than the average size of visible orders on Nasdaq. These results suggest that M-ELO trading is more active in less liquid stocks when the market participants wish to trade large orders. This might seem unrealistic that such a small order type can impact any aspect of markets' behavior. However, in the next section, we show a steady coupling between the share of M-ELO trading and mini-flash crashes' intensity and liquidity provision quality.

1.3 Empirical Approach and Results

This section analyzes the effect of M-ELO trading on market stability which is approximated by the number of mini-flash crashes in individual securities. We also study the effect on various crash characteristics and liquidity measures. Our empirical approach involves relating market stability and liquidity characteristics to the M-ELO trading via stock-week panel regressions. For the panel regressions, we take two methods: (i) two-stage least squares (2SLS) instrumental variable regressions, and (ii) two-stage GMM estimations, which are efficient in the presence of heteroskedasticity of unknown type and apply heteroskedasticity and autocorrelation robust standard errors.

1.3.1 Mini-flash crashes

This paper tries to identify the impact of M-ELO trading on the general number of mini-flash crashes and their characteristics. To account properly for both the cross-section variation and

time variation, we employ the panel structure of the data and estimate the following panel regression with time and fixed effects:

$$y_{it} = \alpha_t + \beta_1 \cdot \text{M-ELO}_{it} + \beta_2 \cdot d_t \cdot \text{M-ELO}_{it} + \theta \cdot X_{it} + C_i + u_{it}, \quad (1.9)$$

where y_{it} is one of the following: (i) weekly number of mini-flash crashes in security i on week t , or (ii) one of the various crash characteristics like Z -score of the extreme crash return, duration of the crash, subsequent price reversal, and others, or (iii) one of the liquidity measures. M-ELO_{it} is a fraction of M-ELO shares matched by Nasdaq to the total number of shares the exchange matched for security i on week t , d_t is the indicator that the holding period of M-ELO orders has been reduced from 500 milliseconds to 10 milliseconds. It equals one for all the days after May 11, 2020, and zero otherwise. X_{it} is a set of control variables that includes stock trading activity characteristics like the average daily number of trades, shares, and volume. It also includes the coefficient of variation as the standard deviation of extreme returns scaled by the mean, the measure of low latency trading from Equation (1.8), and OIBNUM as the measure of order imbalance.

Identifying the causal effect of dark M-ELO trading is generally problematic due to endogeneity. The possibility of reverse causality arises because M-ELO activity may affect market stability, but, at the same time, less stable markets may push participants to the dark. Econometrically, this means that endogenous regressors will make the estimates biased, inefficient, and inconsistent.

One potential solution to the endogeneity problem is the instrumental variable approach in the spirit of Hasbrouck and Saar (2013), Degryse et al. (2015), and Comerton-Forde and Putninš (2015). We instrument M-ELO trading in stock i on week t with the average price of this stock on week t computed from the trades records. To get a more efficient estimator, we use one more instrument which is the price rank of the stock i on week t assigned by the SEC. Thus, we extend the linear regression model (1.9) by the following first-stage regression:

$$\text{M-ELO}_{it} = a_t + \pi_1 Z_{1,it} + \pi_2 Z_{2,it} + \gamma X_{it} + C'_i + v_{it}, \quad (1.10)$$

where $Z_{1,it}$ is the average daily price of stock i on week t , $Z_{2,it}$ is its price rank, and X_{it} is the set of control variables included to weaken the instrument exogeneity assumption.

We can estimate the causal effect of M-ELO_{it} on y_{it} in two steps. At the first stage, we regress M-ELO_{it} on $Z_{1,it}$ and $Z_{2,it}$ to obtain predicted values $\widehat{\text{M-ELO}}_{it} = \hat{\alpha}_t + \hat{\pi}_1 Z_{1,it} + \hat{\pi}_2 Z_{2,it} + \hat{\gamma} X_{it}$. Then, at the second stage, we regress y_{it} on $\widehat{\text{M-ELO}}_{it}$ to obtain the Two Stage Least Squares estimators $\hat{\beta}_{1,2SLS}$ and $\hat{\beta}_{2,2SLS}$.

We believe the selected instruments for the level of M-ELO trading make economic sense because high prices make M-ELO orders less attractive to traders. The higher the price of the security is, the lower the relative bid-ask spread is which makes the price improvement of the M-ELO lower. The same intuition can be applied to the price rank of the security. On the other hand, the price of a security alone is not expected to affect its market stability so as to invalidate exogeneity. The formal tests of instruments' relevance and exogeneity can be found in Table A.1 of the Appendix 1.A.3.

An additional way to handle the endogeneity problem is the construction of a weekly panel of stocks and estimation of a dynamic model using the GMM system estimator developed by Blundell and S. Bond (1998). The property of GMM of not relying on any specific assumption of the distribution of the residuals makes it appropriate for our estimation. To mitigate the bias caused by endogenous regressors, the GMM estimation allows using lagged explanatory variables to eliminate correlations between explanatory variables and error terms. Under these conditions, the resulting estimator consistently estimates the impact of an exogenous change in M-ELO trading activity on the market stability of the stock.

The general form of our dynamic panel regression is as follows:

$$y_{it} = \alpha_t + \beta_1 y_{it-1} + \beta_2 \text{M-ELO}_{it} + \beta_3 \cdot d_t \cdot \text{M-ELO}_{it} + \theta X_{it} + C_i + u_{it}. \quad (1.11)$$

To construct the set of moment conditions we assume sequential exogeneity. As GMM instruments for the lags of the dependent variable, we use its next three further lags. The GMM estimates are robust but typically weakened if the number of instruments is large. This is a common practice to either collapse the instruments to avoid the bias that arises as the number of instruments becomes high or to just use the most recent lags of the dependent variable as instruments.

The regression results for the models (1.9) and (1.11) are reported in Table 1.3. For the panel GMM, we estimate the model (1.11) including the first lag of the dependent variable

(Column 2). The estimation results for both models suggest that relative M-ELO trading effect is highly significant and is negatively associated with mini-flash crash occurrences during the week. Thus, in the linear panel specification (1), the loading on the M-ELO volume relative to the total Nasdaq matched volume indicates that with a one percentage point increase in the relative volume of M-ELO orders, the average number of mini-flash crashes for that week decreases by about 0.14 in that specific security for the period from March 2018 to May 2020.

In contrast, the reduction of the holding period from 500 milliseconds to 10 milliseconds resulted in a weakening of the initial market-stabilizing effect of M-ELO trading. After May 11, 2020, one percentage point increase in relative M-ELO volumes is associated with on average 0.04 decrease in the weekly number of mini-flash crashes. We observe, that the 98% reduction in speed bump properties of M-ELO orders comes together with 71% reduction in the initial stability improvements associated with the M-ELO trading.

Table 1.3 around here

Table 1.3 shows that for the specification (2) of the panel GMM, the first lag of the number of mini-flash crashes turns out to be significant as well. Having all the parameters fixed, an additional mini-flash crash on the previous week is associated with a 0.37 decrease in the average number of crashes on a current week.

The estimation results from the panel GMM model suggests a bigger economic effect of M-ELO trading. With one percentage point increase in relative M-ELO volume during the first two years, the average number of crashes decreased by 0.28. After the design change in May 2020, this effect diminishes by 78.6% and equals 0.06 points of decrease.

We further investigate if M-ELO trading can explain the variation in different crash characteristics. Table 1.4 reports the estimation results of the panel regression where the dependent variables are the extreme return's Z -score at the time of the crash, crash return in absolute value, crash duration, number of trades during the crash, and the relative volatility of extreme returns during the next 30 minutes.

Since we observe 10,113 mini-flash crashes throughout the sample period, and, more specifically, only 5,297 stock-week observations with a non-zero number of crashes, we decide to

proceed with 2SLS linear panel estimation⁴.

Table 1.4 around here

The effect of M-ELO volume relative to the total volume matched by Nasdaq is statistically significant for the absolute value of the mini-flash crash return as well as for its Z -score. The effect is contrasting, however. An increase of one percentage point in M-ELO activity is associated with an increase of 22.7 basis points in the absolute return during the crash, but at the same time with a 0.398 decrease in the Z -score of this return. This result may suggest that for the securities with a higher degree of M-ELO trading flash crashes may stand out less in terms of how volatile they are compared to usual periods. At the same time the size of the crash, in terms of crash returns, tends to be bigger.

There is a significant positive effect of M-ELO on the number of trades that happen during the crash. There are about 135 more trades happen during the mini-flash crash if the security's M-ELO share increases by one percentage point. The model finds no impact of M-ELO neither on crash duration nor on the price reversal in the following 10 minutes.

All the impact of M-ELO trades on crash characteristics is mitigated by approximately 83.3% during the period when the holding time of M-ELO orders was reduced to 10 milliseconds. Clearly, the speed bump effect of M-ELO is dominating in its impact on the characteristics of the crashes.

Table 1.4 also reports the controls which are best in explaining various mini-flash crash characteristics. Firstly, the coefficient of variation of the returns, or simply our measure of stock return volatility, is significant at the 1% level in almost every model specification. The estimates suggest that more volatile stocks have usually more explosive crash returns, shorter crash duration with a greater number of trades. Secondly, the average daily turnover in terms of trades is usually associated with less extreme crashes, after which the price recovers to a larger extent. However, those effects are barely noticeable economically.

⁴ It is known that when the number of instruments in the GMM setting is increasing the bias in estimates increases as well. That is why OLS regression is likely to have higher statistical power. As it was established in Table 1.3, the effect of M-ELO trading is more sound for the panel GMM model specification. This means that a static model just underestimates the effect at worst.

1.3.2 Liquidity provision measures

We further analyze the impact of relative M-ELO trading on various measures of liquidity. Table 1.5 reports the estimation results of model (1.9), where the dependent variable is the quoted spread, the 5-minute realized spread, a fraction of depth available 10 basis points around the midquote to the total daily volume, and the absolute depth imbalance 10 basis points around the midquote. As spreads, depth, and price impact are high-frequency liquidity measures, we do not expect to find a strong seasonality in those measures at weekly time frames. Therefore, a static panel regression is preferable for the analysis since the estimates have higher statistical power.

Table 1.5 around here

The relative M-ELO traded volume has a “positive” impact on the quoted half spread. Thus, a one percentage point increase in M-ELO trading during the first two years of operation is associated with an increase of 1.33 basis points in the quoted spread. This effect is sizeable and significant at a 1% level. At the same time, more M-ELO activity is accompanied by the increase in the depth available 10 basis points from the midquote. For each percentage point increase in M-ELO trading, there is a 0.55 basis points increase in the fraction of the daily volumes which are available close to the midquote during the trading day.

The positive effect from the increase in the available depth can be undermined by the fact that the absolute depth imbalance also increased due to M-ELO. We observe a 4.2 points increase in the absolute depth imbalance 10 basis points around the midquote for each percentage point increase in relative M-ELO volume. It is worth noting, that the increase in depth imbalance by 4.2 doesn’t mean the selling pressure increases four times.

Similarly to the crash characteristics, after the holding period reduction, a big portion of the impact of M-ELO on liquidity is gone. There is no significant effect of M-ELO on the 5-minute realized spread measure. Table 1.5 also reports estimation results for control variables in each model specification. We observe some significant variables that impact the liquidity provision, but most of the effects are economically negligible.

1.3.3 Robustness

In the following, we detail additional robustness tests to support previous results. As the first robustness exercise, we estimate the model using alternative specifications of M-ELO trading. Previously, we used to relate M-ELO volumes to Nasdaq’s total matched volumes for the main analysis. But M-ELO activity can also be compared to the overall dark volume handled by Nasdaq. This allows to distinguish M-ELO trading from other dark trading activity and to determine any additional or specific impact of M-ELO on market stability measures.

Also, we relate M-ELO volumes to the consolidated volume traded in each particular security. This specification of M-ELO activity takes into account the fact that M-ELO orders are available only to Nasdaq’s participants, while Nasdaq may not have the biggest share in trading for some particular stock. Table 1.6 reports the results of the analysis for M-ELO volume related to Nasdaq’s dark volume in columns 1 and 3, and total consolidated volume in columns 2 and 4. The estimation results for both static and dynamic panels suggest the effect of M-ELO stays highly significant and becomes more pronounced economically.

Table 1.6 around here

As an additional robustness test, we estimate the model separately for big and small stocks. We define a stock as a big one if its daily dollar trading volume is above all stocks’ median trading volume throughout the sample period. Columns 1 and 2 in Table 1.7 report the estimation results for samples of big and small stocks separately. The estimates suggest that the results are driven by the most actively traded securities. There is no significant effect of M-ELO orders on the stability of small stocks, while the effect on the sample of big stocks is similar to what we observe in the full sample.

Table 1.7 around here

Finally, in column 3 of Table 1.7, we report the estimation results of the linear panel model in Equation (1.9) after we remove the outliers in the controls. The estimates indicate a very similar effect of the M-ELO activity on the market stability. For the period from March 2018 to May 2020 a higher degree of M-ELO trading was associated with fewer number of mini-flash crashes, while, after the holding period of M-ELO orders was decreased by 98%, this positive effect reduced by 79.9%.

1.4 Conclusion

This paper provides novel evidence on market stability and liquidity provision due to the implementation of a non-displayed (dark) Midpoint Extended Life Order (M-ELO). M-ELO is a dark order that cannot interact with lit (visible) orders. It also possesses the speed bump effect due to the holding period prior to the execution. We use high-frequency order book message data from the Nasdaq exchange for the three years of M-ELO existence. The rule change applied on May 11, 2020, makes it possible to disentangle the dark and the speed bump impacts of M-ELO orders on market stability and liquidity.

For the period from January 2018 to October 2020, the degree of M-ELO activity is associated with a lower frequency of mini-flash crashes for Nasdaq traded securities. Results from panel regressions suggest the presence of significant effects of the M-ELO trading on crash returns, volatility, and trading activity. Higher relative volumes traded via M-ELO are associated with less turbulent crashes, which is more desirable for long-term investors. The effect of the M-ELO on the quality of the liquidity provision is mixed. We document the increase of both quoted spread and depth close to the midquote due to the M-ELO trading. At the same time, increased order imbalance may undermine the positive effect of the improved market depth.

Our analysis shows that trading activity in M-ELO impacts market stability and liquidity mainly due to the speed bump effect. The reduction in the M-ELO's holding period by 98% decreases the influence of M-ELO on the market by 80% on average. The robustness of the results to different specifications of the model strengthens the conclusion that only about 20% of the M-ELO market stabilizing effect comes from its dark properties and 80% from the speed bump properties.

As M-ELO volumes are relatively small, we are cautious about extrapolating the results of this analysis. The main goal of our research is to identify the effects of M-ELO on market stability during recent years. Our study delivers an important insight for market participants, policymakers, and researchers. The trade-off between the execution speed and order transparency is capable of impacting the general stability of financial markets.

References

- Andersen, Torben G, Tim Bollerslev, Francis X Diebold, and Heiko Ebens (2001). “The distribution of realized stock return volatility”. *Journal of Financial Economics* 61.1, pp. 43–76.
- Bellia, Mario, Kim Christensen, Aleksey Kolokolov, Loriana Pelizzon, and Roberto Renò (2020). “High-frequency trading during flash crashes: walk of fame or hall of shame?” *Unpublished working paper*.
- Belter, Klaus (2007). “Supply and information content of order book depth: the case of displayed and hidden depth”. *Unpublished working paper*.
- Biais, Bruno and Thierry Foucault (2014). “HFT and market quality”. *Bankers, Markets & Investors* 128, pp. 5–19.
- Biais, Bruno, Thierry Foucault, and Sophie Moinas (2015). “Equilibrium fast trading”. *Journal of Financial Economics* 116.2, pp. 292–313.
- Blundell, Richard and Stephen Bond (1998). “Initial conditions and moment restrictions in dynamic panel data models”. *Journal of Econometrics* 87.1, pp. 115–143.
- Boulatov, Alex and Thomas J. George (2013). “Hidden and displayed liquidity in securities markets with informed liquidity providers”. *Review of Financial Studies* 26.8.
- Brogaard, Jonathan, Allen Carrion, Thibaut Moyaert, Ryan Riordan, Andriy Shkilko, and Konstantin Sokolov (2018). “High frequency trading and extreme price movements”. *Journal of Financial Economics* 128.2, pp. 253–265.
- Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan (2014). “High-frequency trading and price discovery”. *Review of Financial Studies* 27.8, pp. 2267–2306.
- Buti, Sabrina, Barbara Rindi, and Ingrid M. Werner (2017). “Dark pool trading strategies, market quality and welfare”. *Journal of Financial Economics* 124.2, pp. 244–265.
- Carrion, Allen (2013). “Very fast money: High-frequency trading on the NASDAQ”. *Journal of Financial Markets* 16.4, pp. 680–711.
- Chordia, Tarun, Richard Roll, and Avanidhar Subrahmanyam (2002). “Order imbalance, liquidity, and market returns”. *Journal of Financial Economics* 65.1, pp. 111–130.

- Comerton-Forde, Carole, Katya Malinova, and Andreas Park (2018). “Regulating dark trading: Order flow segmentation and market quality”. *Journal of Financial Economics* 130.2, pp. 347–366.
- Comerton-Forde, Carole and Talis J. Putninš (2015). “Dark trading and price discovery”. *Journal of Financial Economics* 118.1, pp. 70–92.
- Degryse, Hans, Frank De Jong, and Vincent van Kervel (2015). “The impact of dark trading and visible fragmentation on market quality”. *Review of Finance* 19.4, pp. 1587–1622.
- Easley, David, Marcos López de Prado, and Maureen O’Hara (2012). “Flow toxicity and liquidity in a high-frequency world”. *Review of Financial Studies* 25.5, pp. 1457–1493.
- Foley, Sean and Talis J. Putninš (2016). “Should we be afraid of the dark? Dark trading and market quality”. *Journal of Financial Economics* 122.3, pp. 456–481.
- Foucault, Thierry, Marco Pagano, Ailsa Roell, and Ailsa Röell (2013a). *Market liquidity: theory, evidence, and policy*. Oxford University Press.
- Golub, Anton, John Keane, and Ser-Huang Poon (2012). “High frequency trading and mini flash crashes”. *Unpublished working paper*.
- Hasbrouck, Joel and Gideon Saar (2013). “Low-latency trading”. *Journal of Financial Markets* 16.4, pp. 646–679.
- Hu, Edwin (2018). “Intentional Access Delays, Market Quality, and Price Discovery: Evidence from IEX Becoming an Exchange”. *Unpublished working paper*.
- Johnson, Neil, Guannan Zhao, Eric Hunsader, Hong Qi, Nicholas Johnson, Jing Meng, and Brian Tivnan (2013). “Abrupt rise of new machine ecology beyond human response time”. *Scientific Reports* 3, p. 2627.
- Kirilenko, Andrei, Albert S Kyle, Mehrdad Samadi, and Tugkan Tuzun (2017). “The Flash Crash: High-frequency trading in an electronic market”. *Journal of Finance* 72.3.
- Leal, Sandrine Jacob, Mauro Napoletano, Andrea Roventini, and Giorgio Fagiolo (2016). “Rock around the clock: An agent-based model of low- and high-frequency trading”. *Journal of Evolutionary Economics* 26.1, pp. 49–76.
- O’Hara, Maureen, Chen Yao, and Mao Ye (2014). “What’s not there: Odd lots and market data”. *Journal of Finance* 69.5, pp. 2199–2236.
- Zhu, Haoxiang (2014). “Do dark pools harm price discovery?” *Review of Financial Studies* 27.3, pp. 747–789.

Tables and Figures

Table 1.1. Descriptive statistics of the sample firms.

The data set covers observations for 196 firms and exchange traded funds (ETFs) for the period from January 22, 2018 to October 2, 2020. The table shows the mean, standard deviation, and quartiles of all variables. The coefficient of variation of the extreme returns (σ_r/μ_r) shows the extent of return variability in relation to its mean, where extreme returns are maximum possible returns during crash identification intervals. Nasdaq's share denotes the share of consolidated traded volume handled by Nasdaq. QS is quoted half spread, RS_{5min} and PI_{5min} are realized spread and price impact in the following 5 minutes after the trade, respectively. Depth(10) is the U.S. dollar value of shares available 10 basis points around the midquote. DI(10) shows the imbalance of buy-sell orders 10 basis points around the midquote. OIBNUM represents the order imbalance in terms of trades. The strategic runs variable shows the fraction of time, high-frequency traders (HFTs) engage in strategical order submission during the day. Msg/Trades represents the ratio of all order modification messages relative to executed trades. Msg/\$100 shows how many order add messages are submitted for every \$100 traded. The statistics are equally weighted based on a daily observations per firm.

	Mean	StDev	25th	50th	75th
General Characteristics					
Price	110.11	203.41	40.61	62.08	113.65
Trades '000	16.05	24.40	6.29	9.25	15.74
Shares, 'M	1.99	3.10	0.58	0.99	2.05
Volume \$'M	201.64	593.86	35.13	68.56	140.44
σ_r/μ_r	0.63	0.17	0.52	0.59	0.69
Nasdaq's share	0.23	0.11	0.13	0.19	0.33
Liquidity Measures					
QS, bps	4.59	6.17	2.08	2.99	4.73
RS_{5min} , bps	-0.19	6.11	-1.56	-0.22	1.09
PI_{5min} , bps	0.07	6.27	-1.18	0.08	1.36
Depth(10), \$'M	1.55	3.76	0.19	0.41	1.04
DI(10)	3.42	35.57	-0.38	2.66	7.26
OIBNUM, %	-1.09	10.49	-7.30	-1.03	5.11
Algorithmic Trading (AT) Measures					
Strategic Runs, %	1.83	7.03	0.13	0.36	1.02
Msg/Trades	39.96	25.88	23.82	33.46	48.27
Msg/\$100	0.69	0.63	0.30	0.52	0.86

Table 1.2. Descriptive statistics of the identified mini-flash crashes.

The data set covers observations for 196 firms and exchange traded funds (ETFs) for the period from January 22, 2018 to October 2, 2020. The table shows the mean, standard deviation, and quartiles of various crash characteristics for 10,113 identified mini-flash crashes. The Z-score of the return is the value of $\frac{r_i - \mu_r}{\sigma_r}$, where r_i is the extreme return on the interval i , μ_r and σ_r are the mean and standard deviation of extreme returns on that day. The reversal shows what fraction of the initial jump did the price retrace 10 minutes after the crash. Number of trades, shares traded, and dollar volume traded are counted during the period of the crash. Aftermath volatility shows the size of the standard deviation of the extreme returns during the next 30 minutes after the crash ends relative to the standard deviation of the extreme returns throughout the day. M-ELO/Matched represents the relative amount of shares traded with M-ELO orders compared to the total amount of shares matched by the Nasdaq exchange. The statistics are equally weighted based on a daily observations per firm.

	Mean	StDev	25th	50th	75th
Return, bps	-0.08	110.70	-60.15	-10.98	56.30
Abs. Return, bps	77.99	78.56	35.58	58.33	95.51
Return Z-score	8.53	1.55	7.48	8.07	9.04
Duration, s	47.91	37.43	19.00	38.00	68.00
Reversal	0.78	0.44	0.47	0.74	1.00
# of Trades	305.38	246.83	151.00	248.00	393.38
# of Shares, '000	46.46	73.96	12.66	25.43	51.06
Volume, \$'M	3.94	5.85	0.99	1.99	4.50
Aftermath Volatility	1.16	0.42	0.92	1.10	1.30
M-ELO/Matched, %	1.98	3.08	0.26	0.83	2.43

Table 1.3. The effect of M-ELO trading on the average weekly number of mini-flash crashes. The table reports the estimation results for the following regressions (with and without a lag of $y_{i,t}$):

$$y_{it} = \alpha_t + \beta_1 y_{i,t-1} + \beta_2 \text{M-ELO}_{it} + \beta_3 \cdot d_t \cdot \text{M-ELO}_{it} + \theta X_{it} + C_i + u_{it},$$

which is estimated on a sample of 196 liquid stocks traded on Nasdaq from January 22, 2018 to October 2, 2020. The specification in column (1) is a linear static panel instrumental variables model with time and fixed effects. The model specification in column (2) describes a dynamic panel estimated using GMM with the three most recent lags of the dependent variable as GMM instruments. The dependent variable is the average number of mini-flash crashes. The set of control variables includes a coefficient of variation of extreme returns, daily number of trades, number of shares, and U.S. dollar volume traded, the AT measure of Hasbrouck and Saar (2013), and the OIBNUM measure of order imbalance. This is an unbalanced panel with weekly observations. M-ELO is the share of the Midpoint Extended Life Order volume relative to the total volume handled by Nasdaq for the particular stock, d_t is a dummy variable that equals one at the time when Nasdaq decreased the M-ELO holding period by 98%, and zero otherwise. The value of M-ELO trading is instrumented by its daily average price and the price rank assigned by the SEC. Heteroskedasticity corrected t -statistics are reported in parentheses. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

	<i>Dependent variable:</i>	
	Average weekly number of crashes	
	<i>Panel linear</i>	<i>Panel GMM</i>
	(1)	(2)
$y_{i,t-1}$		-0.367*** (-12.396)
M-ELO	-14.356*** (-3.400)	-27.622** (-2.292)
$d_t \cdot \text{M-ELO}$	10.754*** (3.117)	21.269** (2.032)
Controls	Yes	Yes
Observations	16,116	12,199
F-statistic	5,598.2***	1,133.2***

Table 1.4. The effect of M-ELO trading on crash characteristics.

The table reports the estimation results for the following linear panel regression:

$$y_{it} = \alpha_t + \beta_1 \text{M-ELO}_{it} + \beta_2 \cdot d_t \cdot \text{M-ELO}_{it} + \theta X_{it} + C_i + u_{it},$$

which is estimated on a sample of 196 liquid stocks traded on Nasdaq from January 22, 2018 to October 2, 2020. The dependent variables are extreme return Z-score during a crash, the absolute value of the crash return, crash duration in seconds, the number of trades executed during the crash and the price reversal 10 minutes after the crash. M-ELO is the share of the Midpoint Extended Life Order volume relative to the total volume handled by Nasdaq for the particular stock, d_t is a dummy variable that equals one at the time when Nasdaq decreased the M-ELO holding period by 98%, and zero otherwise. The value of M-ELO trading is instrumented by the stock daily average price and the price rank assigned by the SEC. The control variables are the daily number of trades, shares traded and dollar volumes. σ_r/μ_r is the coefficient of variation for extreme returns during the day. AT is a measure of algorithmic trading defined in Equation (1.8), OIBNUM is a measure of order imbalance based on number of buy and sell transactions. This is an unbalanced panel with weekly observations. Heteroskedasticity corrected t -statistics are reported in parentheses. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

	<i>Dependent variable:</i>				
	Z-score (1)	Return (2)	Duration (3)	# Trades (4)	Reversal (5)
M-ELO	-39.814** (-2.164)	0.227* (1.823)	326.575 (0.684)	$1.354 \cdot 10^4$ * (1.909)	0.478 (0.056)
$d_t \cdot \text{M-ELO}$	31.651** (2.134)	-0.185* (-1.803)	-261.458 (-0.680)	$-1.2 \cdot 10^4$ ** (-2.076)	-0.179 (-0.026)
Trades '000	0.004 (1.195)	$-2.373 \cdot 10^{-5}$ ** (-2.081)	-0.185** (-2.548)	1.608** (2.265)	0.003*** (3.643)
Volumes '\$B	-0.150 (-1.297)	$-3.476 \cdot 10^{-4}$ (-1.112)	6.806** (2.083)	-25.337 (-1.278)	-0.032 (-0.653)
Shares 'M	-0.011 (-0.522)	$2.72 \cdot 10^{-6}$ (0.035)	-0.925*** (-2.808)	-1.366 (-0.241)	0.005 (1.221)
σ_r/μ_r	3.514*** (14.609)	0.013*** (8.241)	-22.417*** (-5.106)	357.209*** (6.450)	0.057 (0.870)
AT	0.507 (1.589)	-0.002 (-1.020)	1.817 (0.275)	-325.483*** (-2.682)	0.153 (1.502)
OIBNUM	-0.239 (-0.972)	$-9.049 \cdot 10^{-4}$ (-0.893)	-3.520 (-0.620)	-176.279** (-2.443)	0.085 (1.169)
Observations	5,297	5,297	5,297	5,297	5,297
R ²	0.073	0.066	0.034	0.009	0.047
F-statistic	676.672***	792.883***	224.239***	368.361***	255.157***

Table 1.5. The effect of M-ELO trading on market liquidity.
The table reports the estimation results for the following linear panel regression:

$$y_{it} = \alpha_t + \beta_1 \text{M-ELO}_{it} + \beta_2 \cdot d_t \cdot \text{M-ELO}_{it} + \theta X_{it} + C_i + u_{it},$$

which is estimated on a sample of 196 liquid stocks traded on Nasdaq from January 22, 2018 to October 2, 2020. Dependent variables are quoted half-spread, 5-minute realized spread, dollar depth available 10 basis points around the midquote relative to the average daily dollar trading volume, and depth imbalance 10 basis points around the midquote. M-ELO is the share of the Midpoint Extended Life Order volume relative to the total volume handled by Nasdaq for the particular stock, d_t is a dummy variable that equals one at the time when Nasdaq decreased the M-ELO holding period by 98%, and zero otherwise. The value of M-ELO trading is instrumented by the stock daily average price and the price rank assigned by the SEC. Trades, volumes, and shares are daily averages, σ_r/μ_r is the coefficient of variation for extreme returns during the day. AT is a measure of algorithmic trading defined in Equation (1.8), OIBNUM is a measure of order imbalance based on number of buy and sell transactions. This is an unbalanced panel with weekly observations. Heteroskedasticity corrected t -statistics are reported in parentheses. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

	<i>Dependent variable:</i>			
	QS (1)	RS _{5min} (2)	$\frac{\text{Depth}(10)}{\$ \text{Volumes}}$ (3)	DI(10) (4)
M-ELO	133.317** (2.281)	34.570 (0.910)	0.554*** (2.881)	419.619*** (2.967)
$d_t \cdot \text{M-ELO}$	-111.088** (-2.244)	-26.939 (-0.840)	-0.496*** (-2.999)	-316.960*** (-2.718)
Trades '000	0.023* (1.713)	0.029** (2.457)	$3.219 \cdot 10^{-4}$ ** (2.542)	-0.023 (-0.946)
Volumes '\$B	-1.712** (-1.970)	-0.775* (-1.779)	-0.006 (-1.416)	-3.953*** (-3.134)
Shares 'M	0.414*** (4.012)	-0.098 (-1.612)	-0.003*** (-3.370)	1.048*** (4.829)
σ_r/μ_r	2.727*** (4.314)	0.130 (0.167)	-0.016*** (-4.377)	2.705** (2.216)
AT	-8.201** (-2.559)	1.015 (0.518)	-0.004 (-0.569)	-7.600** (-2.157)
OIBNUM	0.615 (1.400)	-0.260 (-0.564)	$-8.253 \cdot 10^{-5}$ (-0.038)	2.447* (1.902)
Observations	16,116	16,096	16,116	16,080
R ²	0.008	$4.196 \cdot 10^{-4}$	0.030	0.006
F-statistic	544.890***	23.635***	786.298***	348.477***

Table 1.6. Alternative M-ELO activity specifications.

The table reports the estimation results for the following regressions (with and without a lag of $y_{i,t}$):

$$y_{it} = \alpha_t + \beta_1 y_{i,t-1} + \beta_2 \text{M-ELO}_{it} + \beta_3 \cdot d_t \cdot \text{M-ELO}_{it} + \theta X_{it} + C_i + u_{it},$$

which are estimated on a sample of 196 liquid stocks traded on Nasdaq from January 22, 2018 to October 2, 2020. The specification in columns (1) and (2) is a linear static panel instrumental variables model with time and fixed effects. The specifications for columns (3) and (4) describe a dynamic panel estimated using GMM with six most recent lags of the dependent variable as GMM instruments for differenced equation. The dependent variable is the average number of mini-flash crashes. M-ELO (Dark) and M-ELO (Cons) represent a share of M-ELO volume relative to, respectively, the volume of dark trading on Nasdaq, and total consolidated volume across exchanges, d_t is a dummy variable that equals one at the time when Nasdaq decreased the M-ELO holding period by 98%, and zero otherwise. The set of control variables includes a coefficient of variation of extreme returns, average daily number of trades, number of shares, U.S. dollar volume traded, the AT measure of Hasbrouck and Saar (2013), and the OIBNUM measure of order imbalance. This is an unbalanced panel with weekly observations. The value of M-ELO trading is instrumented by its daily average price and the price rank assigned by the SEC. Heteroskedasticity corrected t -statistics are reported in parentheses. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

	<i>Dependent variable:</i>			
	Average weekly number of crashes			
	<i>Panel linear</i>		<i>Panel GMM</i>	
	(1)	(2)	(3)	(4)
$y_{i,t-1}$			-0.323*** (-12.079)	-0.320*** (-12.154)
M-ELO (Dark)	-17.383*** (-2.693)		-36.608*** (-2.941)	
M-ELO (Cons)		-89.692*** (-3.138)		-156.464*** (-3.075)
$d_t \cdot \text{M-ELO (Dark)}$	14.860*** (2.633)		31.597*** (2.793)	
$d_t \cdot \text{M-ELO (Cons)}$		71.634*** (2.901)		-71.514 (-1.092)
Controls	Yes	Yes	Yes	Yes
Observations	16,116	16,116	12,199	12,199
F-statistic	5,754.6***	5,650.7***	1,161.2***	1,192.6***

Table 1.7. The effect of M-ELO trading on the number of mini-flash crashes for big and small stocks and outlier-robust effects.

This table reports the estimation results for the following linear panel regression:

$$y_{it} = \alpha_t + \beta_1 \text{M-ELO}_{it} + \beta_2 \cdot d_t \cdot \text{M-ELO}_{it} + \theta X_{it} + C_i + u_{it},$$

which is estimated on a sample of 196 liquid stocks traded on Nasdaq from January 22, 2018 to October 2, 2020. The dependent variable is the average number of mini-flash crashes. Column (1) reports results for the sub-sample of small stocks, and column (2) for the sub-sample of big stocks in terms of average daily volumes. Column (3) reports results for the full sample but after discarding the highest and the lowest 5% values of the average daily number of trades, shares and U.S. dollar volume. M-ELO is the share of the Midpoint Extended Life Order volume relative to the total volume handled by Nasdaq for the particular stock, d_t is a dummy variable that equals one at the time when Nasdaq decreased the M-ELO holding period by 98%, and zero otherwise. The value of M-ELO trading is instrumented by the stock daily average price and the price rank assigned by the SEC. The set of control variables includes a coefficient of variation of extreme returns, average daily number of trades, number of shares, U.S. dollar volume traded, the AT measure of Hasbrouck and Saar (2013), and the OIBNUM measure of order imbalance. This is an unbalanced panel with weekly observations. Heteroskedasticity corrected t -statistics are reported in parentheses. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

	<i>Dependent variable:</i>		
	Average weekly number of crashes		
	(1)	(2)	(3)
M-ELO	4.584 (1.387)	-17.334*** (-3.426)	-12.813*** (-3.711)
$d_t \cdot \text{M-ELO}$	-4.443 (-1.521)	11.303*** (2.971)	10.239*** (3.465)
Controls	Yes	Yes	Yes
Sample	Small Stocks	Big Stocks	Full
Outliers Removed	No	No	Yes
Observations	7,507	8,609	11,932
R ²	0.290	0.235	0.191
F-statistic	$3.488 \cdot 10^{3***}$	$2.823 \cdot 10^{3***}$	$3.326 \cdot 10^{3***}$

Figure 1.1. Mini-flash crashes, quoted spread, and book depth throughout the sample.

The first panel of the figure plots the total amount of identified mini-flash crashes across the sample of 196 liquid stocks traded on Nasdaq from January 22, 2018 to October 2, 2020. The second panel shows the median quoted spread (in basis points) across securities. The last panel presents the median aggregate monetary value (in thousands of dollars) of shares offered within 10 basis points around the midquote.

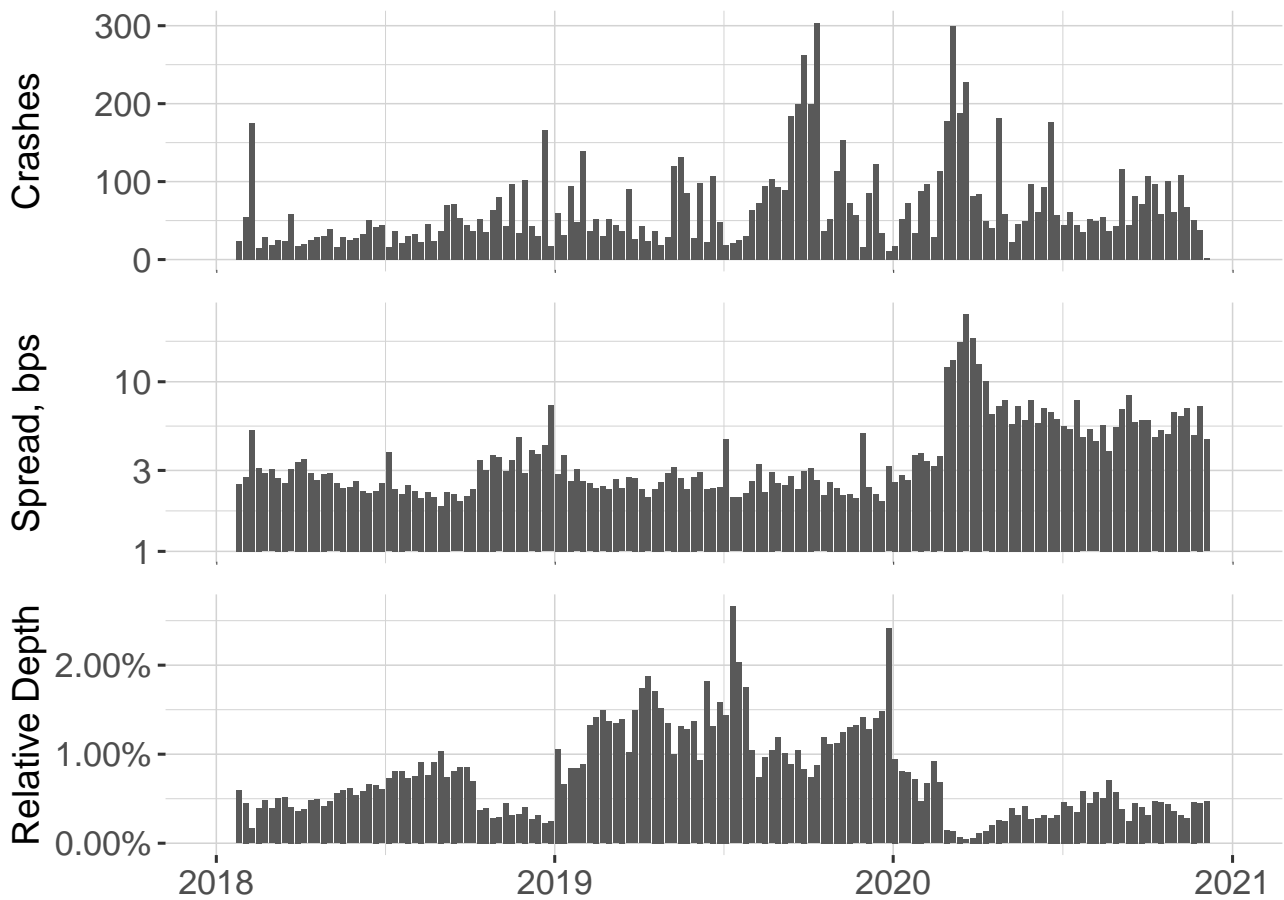


Figure 1.2. Example of the mini-flash crash.

Panel A plots the Procter&Gamble (P&G) share price on March 21, 2018. At 2 p.m., the price experienced a mini-flash crash. The pre-market price of P&G was at around \$78.2, dropped to the region \$77.4 – \$77.7, where it stayed fairly stable until 2 p.m., and experienced then a massive spike to the levels of approximately \$78.3. Within the next five minutes, the price dropped more than 1.4% to \$77.16, and eventually returned to the region of its previous daily consolidation. Panel B zooms in around the time of the crash. Each dot represents a trade. The duration of the crash is 26.2 seconds, the cumulative return of the first spike is 0.98%, the volume traded during the crash is \$633 thousand.

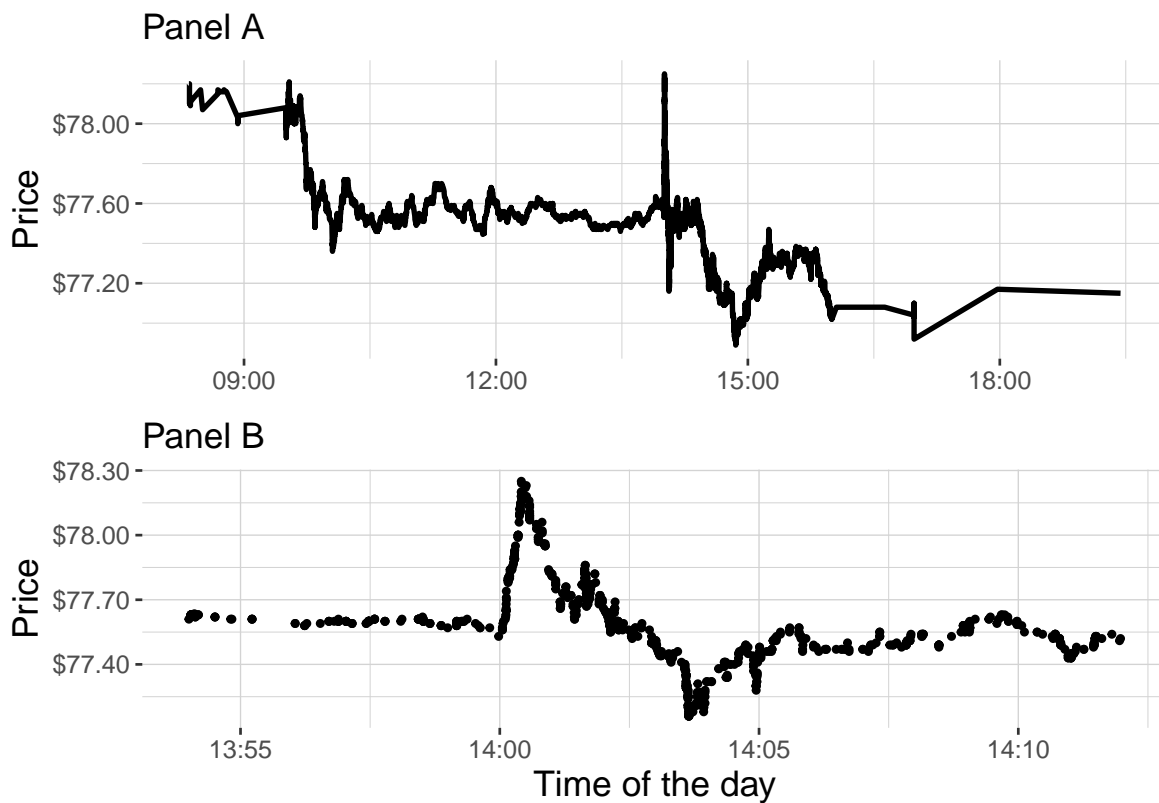


Figure 1.3. Share of M-ELO executions.

This figure plots time trends in M-ELO volumes relative to the total volumes matched by the Nasdaq exchange. The 10th, 50th, and 90th percentiles are depicted. M-ELO orders became available on March 12, 2018. The shaded area represents the period starting when Nasdaq decreased the M-ELO “Holding period” from 500 milliseconds to 10 milliseconds.

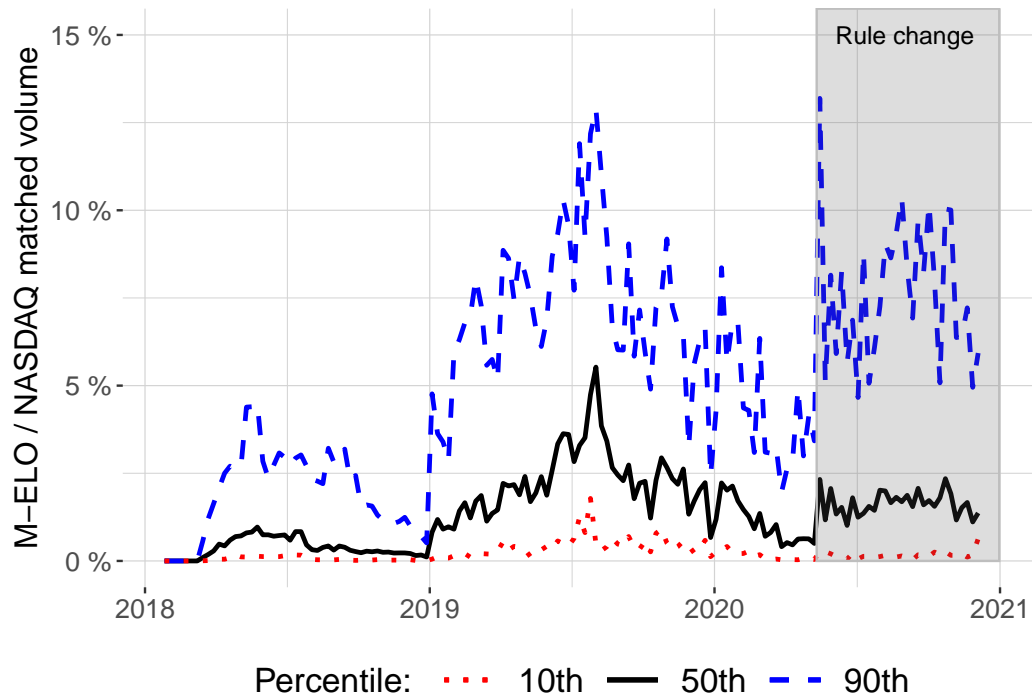
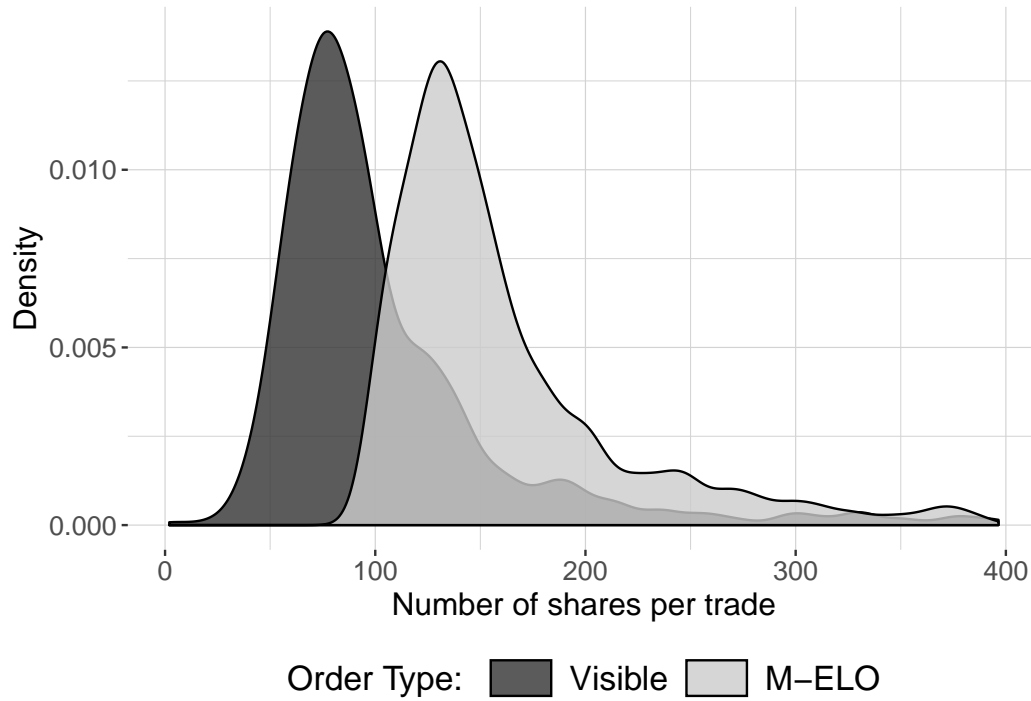


Figure 1.4. Densities of lit (visible) order sizes and M-ELO order sizes.

This figure plots densities of order sizes of two types of orders: (i) visible limit orders, and (ii) non-displayed, M-ELO orders. For better representation, only order sizes less than 400 shares are considered. All observations are stock-day averages.



1.A Appendix

1.A.1 M-ELO order specifications

- Non-displayed order type: M-ELO executions are reported to the Securities Information Processors and provided in Nasdaq's proprietary data feed in the same manner as all other transactions occurring on Nasdaq (i.e., without any new or special indication that a transaction is an M-ELO execution).
- The M-ELO timer for 500 millisecond waiting period starts upon entry if the order is marketable at the midpoint. If the order is not eligible to trade at the midpoint upon the entry, the M-ELO timer will start when the price of the order is at or better than the midpoint of the NBBO.
- Effective May 11, 2020, the holding period was reduced from 500 to 10 milliseconds.
- Any modification on a resting M-ELO order will result in a restart of the timer, except in the case of reducing the order quantity.
- The timer does not reset if the NBBO moves.
- An M-ELO order is ranked in time priority among other M-ELO orders at the time it becomes eligible to execute.
- Only round lots are accepted for the M-ELO submission.
- M-ELO orders may execute in a locked market but not in a crossed market⁵.
- The M-ELO order type will never route out.

⁵In a locked market, a stock's bid price at one exchange and ask price at another exchange are identical, that is, the bid-ask spread is zero. In a crossed market, the bid price exceeds the ask price.

1.A.2 Stocks and ETFs selected for panel regressions

Below, we list Nasdaq tickers of companies and ETFs that meet the minimum liquidity requirement (in terms of trading activity) on every day throughout the sample period:

AABA, AAL, AAPL, ABBV, ABT, ADBE, ADI, ADSK, AMAT, AMD, AMGN, AMTD, AMZN, APA, APC, ATVI, AVGO, AXP, BA, BABA, BAC, BHGE, BIDU, BIIB, BK, C, CAH, CAT, CELG, CHTR, CMCSA, CME, COP, COST, CRM, CSCO, CSX, CTRP, CTSH, CVS, CVX, DAL, DHI, DIS, DISCA, DISH, DLTR, DUK, DVN, DWDP, EA, EBAY, EEM, EFA, EOG, ETFC, EWZ, EXPE, FAST, FB, FCX, FITB, FOXA, FXI, GDX, GE, GILD, GM, GOOG, GOOGL, GPS, GS, HAL, HD, HES, IBM, IEMG, INTC, INTU, IWM, IYR, JD, JNJ, JPM, KHC, KO, KR, KRE, KSS, LB, LBTYK, LEN, LOW, LRCX, LUV, M, MA, MCD, MCHI, MDLZ, MDT, MET, MGM, MNST, MO, MOMO, MPC, MRK, MRO, MRVL, MS, MSFT, MTCH, MU, MXIM, MYL, NFLX, NKE, NKTR, NTAP, NVDA, ON, ORCL, OXY, PCG, PEP, PFE, PG, PM, PTEN, PYPL, QCOM, QQQ, RIG, ROST, SBUX, SCHW, SLB, SMH, SPLK, SPY, SQ, STLD, STX, SWKS, T, TGT, TJX, TLT, TMUS, TQQQ, TSLA, TSM, TTWO, TVIX, TWTR, TXN, UAL, UNH, UNP, UPS, URBN, USB, UTX, UVXY, V, VALE, VLO, VXX, VZ, WB, WBA, WDAY, WDC, WFC, WMB, WMT, WYNN, X, XBI, XEL, XLE, XLF, XLI, XLK, XLNX, XLP, XLU, XLV, XLY, XOM, XOP, XRAY, XRT, YNDX, ZION.

1.A.3 Tests for the instruments

Table A.1. Properties of the instrumental variables.

The table reports the results of instrument relevance and exogeneity tests. To explain the variation in the relative volume of M-ELO orders in the equation (1.9) the following instruments are used: $Z_{1,it}$ is the average price in dollars of the stock i on week t , and $Z_{2,it}$ is the price rank of the stock i on week t assigned by the SEC. The instruments' relevance ($\text{Cov}(Z, \text{M-ELO}) \neq 0$) is tested by estimating the first-stage regression, and obtaining F -statistics resulting from the test $H_0 : \pi_1 = \pi_2 = 0$ against the alternative $H_1 : \pi \neq 0$ or $\pi_2 \neq 0$. The rule of thumb suggests that the F -statistics for joint significance of the instruments in the first-stage should exceed 10. The instrument exogeneity assumption is weakened by including the control variables X_{it} into the first-stage regression. The overidentifying test is implemented by, first, obtaining the residuals of the 2SLS model:

$$\hat{u}_{it}^{2SLS} = y_{it} - \hat{\alpha}_t - \hat{\beta}_1 \text{M-ELO}_{it} - \hat{\beta}_2 \cdot d_t \cdot \text{M-ELO}_{it} - \hat{\theta} X_{it},$$

and then regressing these residuals on the instruments and control variables. The resulting J -statistic of the test $H_0 : \eta_1 = \eta_2 = 0$ versus $H_1 : \eta_1 \neq 0$ or $\eta_2 \neq 0$ is distributed according to χ_q^2 where q is the number of instruments minus the number of endogenous regressors.

Panel A: Instrument Relevance

Regression	$\text{M-ELO}_{it} = a_t + \pi_1 Z_{1,it} + \pi_2 Z_{2,it} + \gamma X_{it} + C_i + v_{it},$
Hypothesis	$H_0 : \pi_1 = \pi_2 = 0$
Statistics	$F(2, 15791) = 77.745$
p -value	$< 2.2 \cdot 10^{-16}$

Panel B: Instrument Exogeneity

Regression	$\hat{u}_{it}^{2SLS} = a_t + \eta_1 Z_{1,it} + \eta_2 Z_{2,it} + \gamma X_{it} + C_i + e_{it}$
Hypothesis	$H_0 : \eta_1 = \eta_2 = 0$
Statistics	$J = mF = 2 \cdot 1.7878 \sim \chi_1^2$
p -value	0.0586

Chapter 2

Dark Pools and Price Discovery in Limit Order Markets

2.1 Introduction

Dark pools are trading venues where the transparency of trading is either low or absent. There exists a variety of different types of dark venues from Alternative Trading Systems (in the USA) and Multilateral Trading Facilities (in Europe) regulated as broker-dealers to exchange-based pools. Figure 2.1 shows that the volume executed on such venues represents a significant share of the consolidated equity volume both for the US and European markets.

Figure 2.1 around here

The proliferation of dark pools raises concerns about their impact on market quality and price discovery. As no market participant is aware of what are others doing in the dark pool, such venues serve as a place to go when a trader wants to disclose as little information as possible. Also, when transacting in the dark, a trader often receives a price improvement but faces non-execution risks. The trade-off between the two is crucial in determining the order flow to dark pools.

Regulators and policymakers would like to know more about the implications of active dark trading, but the current financial market design makes it challenging to model. The majority of trading happens nowadays in limit order markets. In contrast to dealer markets where participants trade only with a designated market maker, in limit order markets traders may

transact with each other directly using limit and market orders. This highly complex and non-linear nature of exchanges forces researchers to resort to numerical solutions. Only recently there appeared models that account for the possibility of dark trading alongside the limit order book based exchange.

The majority of the research that focuses on the question of price discovery is still employing the dealer market structure of the exchange. Ye (2012) studies how an informed trader splits orders between an exchange based on Kyle (1985) model and a dark pool as a crossing network. He finds that the price discovery on the exchange reduces due to the dark trading. Zhu (2014), in turn, assumes that both informed and uninformed traders may freely select venues and concludes that under certain conditions there is an improvement in price discovery. He argues that exchanges are more attractive to informed traders, while dark pools are more attractive to uninformed ones.

Brolley (2020) develops a model where both a limit order book and a dark pool source liquidity from a professional liquidity provider. He finds that investors with high (low) private valuations of the asset migrate to high (low) relative price impact order types. This ranking of traders (“immediacy hierarchy”) predicted by the price improvement plays an important role in determining the market quality and welfare implications of dark pools. However, the dark pool reference price in Brolley (2020) is set by the market maker, while it is worth exploring the implications of dark trading in dynamic market conditions when the reference price is the midpoint of the exchange.

Bayona et al. (2020) try to address this issue and construct a two-period model with asymmetric information where a lit venue competes with a dark pool that derives the execution price from the lit venue. They were able to show that the effects of competition depend on the stock market and trader characteristics and the effects in different periods are not the same, despite investors behaving similarly.

The present paper is closely related to Buti et al. (2017) who in a four-period model study the implications of adding a continuous dark pool to the limit order book. Without considering information asymmetry, the authors show that the introduction of a dark venue that competes with the exchange is associated with on average wider spreads, lower depth, trade creation, and welfare deterioration, especially when the limit order book is illiquid. I improve the model of Buti et al. (2017) by enlarging the strategy space of the traders so they can submit limit orders

at different prices. This allows distinguishing orders with different levels of aggressiveness: from market orders (the most aggressive ones) to aggressive limit orders that undercut the present limit orders, to non-aggressive limit orders. Traders can also refrain from trading or submit an order to the dark pool in case they have an access to it. In contrast to Bayona et al. (2020), the presence of a private valuation for the asset in my model allows for avoiding the introduction of exogenous liquidity trading. All traders remain rational and the distribution of the private valuations controls trading for liquidity needs.

The information asymmetry in the model is realized by the presence of traders who are perfectly informed about the final pay-off of the risky asset. Every trader has a private valuation for the asset that partially determines the optimal strategy of order submission. Nobody observes the activity in the dark, but everybody can bayesian update their beliefs about the current state of the dark pool as well as about the final pay-off of the risky asset.

The equilibrium for this model is determined numerically as the set of order submission probabilities which depend on the asset fundamental volatility, the dispersion of the private valuations for the asset, the relative tick size, dark pool availability, and the amount of informed trading. I allow for fractional values of dark pool availability and informed trading which, in contrast to the previous research, opens the way to identifying the optimal level of dark trading and informed trading instead of just banning dark pools or allowing everybody to use them.

In general, there is more venue competition when dark trading becomes more available. However, I find that there is a strong crowding-out effect in the dark pool. Uninformed investors leave the dark pool rapidly when the fraction of informed trading starts to grow. Interestingly, I find that dark pool trading is very limited in explaining variation in order aggressiveness as well as in execution quality measures of market orders, like bid-ask spread and market depth. On the other hand, I show a decrease in the price discovery due to more active dark trading and better execution quality for limit orders measured by the average fill rates. The effect of a dark pool on aggregate welfare is positive as it allows to realize gains from trade when the limit order book is filled. The effect depends largely on the market and traders' characteristics.

The remainder of this paper is organized as follows. Section 2.2 presents the model of the limit order book and the dark pool. I discuss agents, their strategies, and optimal order submission. In Section 2.3, I demonstrate the procedure of obtaining equilibrium, first, in a single-venue market (dark pool is not available) and, second, in a multiple-venue market. I then

study the venue competition phenomenon in Section 2.4. Section 2.5 discusses the empirical implications of dark pool functioning for price discovery and market quality. Finally, Section 2.6 concludes.

2.2 Model

In this section, I present a model of a Limit Order Book (LOB) with the presence of information asymmetry. Later, I introduce a Dark Pool (DP) to this model to which agents can submit several order types. The orders get executed at the DP immediately, given that there is liquidity available, otherwise they stay in the DP, awaiting execution. I also add the possibility for the order to return to the LOB if the trader specifies her immediacy needs by submitting an Immediate or Cancel (IOC) order. Thus, the DP can deliver price improvement without increasing non-execution risks.

I consider a market with a single risky asset. There are four trading periods and four traders, each trading at only one trading period. The final pay-off of the risky asset, paid at the end of the last trading period, is a random variable V , which takes value “high” (v^H) or “low” (v^L) with equal probabilities. Let $\mu_V = E(V)$ be the expected value of the final payoff, and $\sigma_V = \sqrt{\text{Var}(V)}$ its volatility.

2.2.1 Limit Order Book and Dark Pool

The asset is traded on an exchange that operates through a LOB. The LOB is represented by the prices and quantities of the asset available at those prices. The prices are located on a grid $p = \{p_1, p_2, p_3, p_4\}$, where the individual prices are defined as follows:

$$p_1 = \mu_V - 1.5\tau \tag{2.1}$$

$$p_2 = \mu_V - 0.5\tau \tag{2.2}$$

$$p_3 = \mu_V + 0.5\tau \tag{2.3}$$

$$p_4 = \mu_V + 1.5\tau, \tag{2.4}$$

where τ is the minimal price increment, also called the tick size.

Traders are allowed to trade one share only. Every trader in the model is rational and might either demand liquidity (through market orders) or provide liquidity (through limit orders) at any available level on the price grid. To distinguish limit buy orders from limit sell orders, I use negative values for buy quantities and positive values for sell quantities.

Figure 2.2 illustrates some possible states of the LOB. On Panel A, there are limit orders at all levels of the grid (p_1 and p_2 are occupied by limit buy orders and p_3, p_4 by limit sell orders). Nothing, however, stops traders from adding more liquidity to the LOB. For example, the next trader can submit an additional limit buy order at price level p_2 .

Figure 2.2 around here

Panel B presents the situation when it is not possible to submit a sell limit order at price p_2 . As there is already a buy limit order at price p_3 , placing a sell order at p_2 will result in a negative spread and so-called “crossed market”. I do not allow for the “crossed market” so, even if the trader submits a sell limit order at p_2 , the order will become marketable and will result in immediate execution at a better price (p_3) and will not be any different from a simple market order.

To further simplify the model, I assume the existence of a crowd that will absorb any amount of shares submitted to the price levels far from μ_V . Thus, $q_1 = -\infty$ (quantity at the price level p_1) and $q_4 = \infty$ (at price level p_4). This will reduce the strategy space for traders but still leave them the possibility to undercut their peers by submitting a more aggressive limit order at a higher (lower) price in case they want to buy (sell) the asset. In the end, the state of the LOB is characterized by quantities at price levels p_2 and p_3 , $b_t = [q_{2,t}, q_{3,t}]$.

The dark pool (DP) is an alternative opaque marketplace. It is available to only a fraction α of the traders in the model. No trader can observe neither which orders are submitted to the DP nor how much liquidity is available there. The execution price in the DP is derived from the LOB at each period. The common practice suggests that the execution price is the current middle price on the visible exchange marketplace, defined as the average of the best bid and the best ask prices:

$$p_{\text{Mid},t} = \frac{1}{2} (p_t^{\text{ask}} - p_t^{\text{bid}}) \quad (2.5)$$

For the current model, the best bid (ask) price is defined as the maximum (minimum) price

level at which the negative (positive) number of shares is available.

DP executes orders continuously based on the time priority rule. Dark orders submitted to the DP that were not executed may stay there until the opposite-sided dark order arrives. The trading in the model starts with the empty dark pool in the first period ($DP_1 = 0$). Therefore, agents, after every trading period, update their expectations about the state of the dark pool $DP_t \in \{\dots, -1, 0, 1, \dots\}$ based on the observed activity of previous traders.

If the dark pool is not available, then $DP_t = 0, \forall t$. The aggregate liquidity in the model can be summarized by liquidity in the limit order book and the dark pool $\Omega_t = [b_t, DP_t]$.

2.2.2 Agents and Strategies

Firstly, I distinguish traders based on whether they have access to the DP (with probability α), or do not have access to the DP (with probability $1 - \alpha$). Secondly, as I introduce information asymmetry, the fraction of the traders π will know the realization of the asset's pay-off, while the fraction $1 - \pi$ will know only the distribution of the pay-off (μ_V and σ_V). Throughout the paper, I will call the former “informed” and the latter – “uninformed” traders.

All traders in the model, regardless of their information set and the DP access, are risk-neutral and fully rational. They trade if they can obtain a positive utility from trading. The utility of trader i at period t is defined as follows:

$$\mathcal{U}_{i,t}(\varphi) = (\Pi_t(\varphi, \Omega_t) + \beta_i) \cdot \mathbb{I}_{\text{exec}}, \quad (2.6)$$

where Π_t is a pay-off when implementing strategy φ given the state of the LOB and the expected state of the DP, β_i is the private valuation of the risky asset which can have an arbitrary distribution, \mathbb{I}_{exec} is an indicator function that is equal to one if the execution (trade) happens and zero otherwise.

Only one of four agents trades at each period. The size of the order is limited to one share only. A trader can submit market and limit orders or abstain from trading. If a trader has access to the dark pool, he/she can submit dark orders.

The full list of orders is presented in Table 2.1. All orders can be either buy or sell orders (excluding no trade option). Market orders are executed at the best available price depending on the state of the LOB b_t . Upon submitting a limit order, the trader should specify the limit

price, which can be either p_2 or p_3 since at other price levels the crowd absorbs any amount of shares.

Table 2.1 around here

The first order available for submitting to the DP is a regular dark order. This order is executed at the prevailing midprice in the LOB if the DP contains some resting orders on the opposite side. If there is no counterparty for the limit order at the moment, the order will stay in the DP until it gets eventually executed, or until the trading day ends. During this time, the liquidity profile in the LOB can change and this will affect the midprice. That is why the execution price of the dark order $\varphi_{BD}(\tilde{p}_{Mid,t})$ is a random variable (the same logic applies to $\varphi_{SD}(\tilde{p}_{Mid,t})$).

The other order that may be submitted to the DP is called Immediate or Cancel (IOC) order. If this order cannot be immediately and fully executed on the DP, it will be canceled. Furthermore, it will be routed to the LOB as a market order, where the execution is guaranteed. The usage of routing technologies increases the attractiveness of the DP and thus increases the competition between visible and hidden venues.

2.2.3 Optimal Order Submission

Traders solve the optimization problem and choose an order that maximizes their expected utility of trading. The expected utility varies based on whether the trader is informed or not, and can submit orders to the DP or only to the LOB. It also depends on the observed actions from the previous traders because orders used by them contain information about the final value of the asset. No trading delivers zero utility.

Let us consider utilities of buy orders (utilities of sell orders can be computed using the same logic). The pay-off of a market order contains only final value uncertainty and no execution risks because the trade happens instantly. For the trader i the expected utility from a market order is:

$$E\mathcal{U}_{i,t}(\varphi_{BM}) = E[V|\mathcal{F}_t] - p_t^{\text{ask}} + \beta_i, \tag{2.7}$$

where \mathcal{F}_t is all the information available up to time t about the actions of previous traders.

In contrast to market orders, limit orders introduce non-execution risks. When submitting a limit order, the trader doesn't know whether it will be executed against a future market order.

The trader can increase its execution probability by placing a more aggressive limit order but that comes with a trade-off since the price improvement decreases with order aggressiveness.

$$\mathbb{E}\mathcal{U}_{i,t}(\varphi_{BL}(p_j)) = (\mathbb{E}[V|\mathcal{F}_t] - p_j + \beta_i) \cdot \mathbb{E}(\mathbb{I}_{exec}|\mathcal{F}_t, \Omega_t) \quad (2.8)$$

where \mathbb{I}_{exec} is the indicator function of the execution event. The expected value of this indicator function is the probability of execution for a limit order.

Equation 2.8 shows that the execution probability is conditional on the actions of previous traders as they change the distribution of the asset's pay-off, and on the state of the LOB as it influences the willingness of the next traders to submit market orders versus limit orders.

Dark order executes if there is opposite side liquidity in the DP, or if any next trader becomes a counterparty in the dark:

$$\begin{aligned} \mathbb{E}\mathcal{U}_{i,t}(\varphi_{BD}(\tilde{p}_{Mid,t})) &= (\mathbb{E}[V|\mathcal{F}_t] - p_{Mid,t} + \beta_i) \cdot \Pr(DP_t > 0|\mathcal{F}_t) + \\ &+ \Pr(DP_t \leq 0|\mathcal{F}_t) \cdot \sum_{t^* > t} \left(\mathbb{E}[V - \tilde{p}_{Mid,t^*} + \beta_i|\mathcal{F}_{t^*}] \cdot \right. \\ &\left. \cdot \mathbb{E}(\varphi_{SD} + \varphi_{SIOC}|\mathcal{F}_{t^*}) \cdot \prod_{\substack{k < t^* \\ k > t}} \mathbb{E}(1 - \varphi_{SD} - \varphi_{SIOC}|\mathcal{F}_k) \right), \end{aligned} \quad (2.9)$$

where $\mathbb{E}(\varphi_{SD} + \varphi_{SIOC}|\mathcal{F}_{t^*})$ is the probability that at time t^* a trader submits either sell dark order or sell IOC order, and $\mathbb{E}(1 - \varphi_{SD} - \varphi_{SIOC}|\mathcal{F}_k)$ is the probability that at time k a trader submits none of these orders, but some other order.

Equation 2.9 consists of two parts: the first one represents the fraction of the utility obtained if the dark order executes immediately. The second part takes into account all possibilities of the execution in the next periods. Here the trader forms expectations about the future middle price \tilde{p}_{Mid,t^*} and the strategies of the next agents.

The utility from IOC orders is calculated as follows:

$$\begin{aligned} \mathbb{E}\mathcal{U}_{i,t}(\varphi_{BIOC}(p_{Mid,t})) &= (\mathbb{E}[V|\mathcal{F}_t] - p_{Mid,t} + \beta_i) \cdot \Pr(DP_t > 0|\mathcal{F}_t) \\ &+ \mathcal{U}_{i,t}(\varphi_{BM}) \cdot \Pr(DP_t \leq 0|\mathcal{F}_t) \end{aligned} \quad (2.10)$$

Equation 2.10 shows that, since $p_{Mid,t} < p_t^{ask}$ then $\mathbb{E}\mathcal{U}_{i,t}(\varphi_{BIOC}(p_{Mid,t})) \geq \mathbb{E}\mathcal{U}_{i,t}(\varphi_{BM})$. If a trader has access to the dark pool and believes there is a non-zero probability of having some

positive liquidity in the DP, then he/she would strictly prefer buy IOC order to buy market order. Even if the realization of the DP liquidity does not allow for dark order execution, the trader still accomplishes the execution in the LOB.

Given the utility that each order yields, every trader solves the optimization problem of maximizing its expected utility:

$$\max_{\varphi \in \Phi} E\mathcal{U}_{i,t}(\varphi) \quad (2.11)$$

where $\Phi = \{\varphi_{BM}, \varphi_{SM}, \varphi_{BL}, \varphi_{SL}, \varphi_{BD}, \varphi_{SD}, \varphi_{BIOC}, \varphi_{SIOC}, \varphi_{NT}\}$ ¹ is the set of all strategies available for the trader i at time t .

2.3 Equilibrium

In this section, I discuss the equilibrium in order submitting strategies of traders. Firstly, I will present the single market equilibrium, where the dark pool is not available to any trader. Informed traders, in this case, have no other choice but to submit orders to the LOB and contribute to price discovery. Next, I am going to allow the share α of traders to use the DP and examine the equilibrium in the multiple venue market.

2.3.1 Single-Venue Market Equilibrium

In case the DP is not available, the only venue where the trade may take place is the LOB. Therefore, the set of strategies available for traders is $\Phi \setminus \{\varphi_{BD}, \varphi_{SD}, \varphi_{BIOC}, \varphi_{SIOC}\}$. The equilibrium is represented by order submission probabilities for each trader. Depending on the realization of the private valuation of the asset, a trader may opt for either market order, limit order, or refrain from trading. The equilibrium order submission probabilities depend largely on the distribution of private valuation, as well as on the fundamental volatility of the risky asset and on microstructure frictions like quoted spread.

I solve the model backwards by an iterative process. First of all, I assume that the trades do not contain information about the final pay-off of the asset. This means $\Pr(V = v^H | \mathcal{F}_t) = \frac{1}{2}, \forall t$. The informed traders continue knowing the realization of the asset's pay-off, while the uninformed cannot deduce anything from the observed trades.

¹Buy Limit and Sell Limit orders require traders to specify the limit price. The midprice for dark orders is derived from the LOB.

Then, given all possible states of the LOB the last trader might encounter, I calculate the utility of submitting each order. Based on the private value distribution (parameter β), the asset's final pay-off distribution (parameter σ_V), the spacing of the price grid (parameter τ) and the information set of the trader, I obtain, for each order, the probabilities that it yields the highest utility.

Next, I proceed to the third trader. If the third trader is uninformed, his/her best guess of the asset's pay-off is $EV = \mu_V$. Observing the current state of the LOB, and considering the optimal response of the last trader, I can estimate non-execution risks for a limit order of the third trader. This provides all necessary information to obtain the expected utilities of orders. The same logic is used to unwind the strategies of the second and the first trader.

Finally, I bayesian update the probability that the final asset's pay-off is high ($V = v^H$), given the actions of all traders. For example, after the first trader submitted a buy market order the conditional probability can be computed as follows:

$$\begin{aligned} \Pr(V = v^H | \varphi_{BM}^{(1)}) &= \frac{\Pr(\varphi_{BM}^{(1)} | V = v^H) \cdot \Pr(V = v^H)}{\Pr(\varphi_{BM}^{(1)})} \\ &= \frac{\Pr(\varphi_{BM}^{(1)} | V = v^H)}{\Pr(\varphi_{BM}^{(1)} | V = v^H) + \Pr(\varphi_{BM}^{(1)} | V = v^L)}, \end{aligned} \quad (2.12)$$

where

$$\Pr(\varphi_{BM}^{(1)} | V = v^H) = \pi \cdot \Pr(\varphi_{BM}^{(1)} | V = v^H, \text{inf}) + (1 - \pi) \Pr(\varphi_{BM}^{(1)} | V = v^H, \text{uninf}), \quad (2.13)$$

$$\Pr(\varphi_{BM}^{(1)} | V = v^L) = \pi \cdot \Pr(\varphi_{BM}^{(1)} | V = v^L, \text{inf}) + (1 - \pi) \Pr(\varphi_{BM}^{(1)} | V = v^L, \text{uninf}), \quad (2.14)$$

where $\varphi_{BM}^{(1)}$ is the indicator that the first trader opted for buy market order, *inf/uninf* is the event of the first trader is informed/uninformed.

Similarly, if the second trader implemented a sell limit order at price p_3 after observing a buy market order from the first trader, the conditional probability that the final asset's pay-off is high is obtained as follows:

$$\Pr(V = v^H | \varphi_{SL}^{(2)}(p_3), \varphi_{BM}^{(1)}) = \frac{\Pr(\varphi_{SL}^{(2)}(p_3) | \varphi_{BM}^{(1)}, V = v^H) \cdot \Pr(\varphi_{BM}^{(1)} | V = v^H) \cdot \Pr(V = v^H)}{\Pr(\varphi_{SL}^{(2)}(p_3), \varphi_{BM}^{(1)})} \quad (2.15)$$

If the conditional probabilities change considerably from those I have assumed in the beginning, I update them and solve the model one more time. When the conditional probabilities converge to some stable values, I stop the iterative procedure and declare the model is solved.

The algorithm is summarized in Figure 2.3. The value ϵ represents the threshold change to the actual conditional probabilities, and $\Pr(V = v^H | \mathcal{F}_t)$ is the general indication of the probability that the final pay-off of the asset is v^H given the actions of previous traders. I assume the global convergence of the algorithm but do not rigorously prove it. For the wide range of parameters and considerably small value of ϵ , I obtain the convergence, on average, after six iterations.

Figure 2.3 around here

Figure 2.4 represents part of the tree of the sequential game that the agents play during the trading day². Only strategies with non-zero probabilities in the equilibrium are presented. Consider, for example, the second trader, who observes the empty LOB after the first trader submitted a sell market order. It is not optimal for the second trader to submit aggressive limit orders when the book is empty. At this stage, the trader is better off by submitting a buy limit order at p_2 (rather than at p_3). The trader faces higher non-execution risks (as next traders can undercut him/her by placing a buy limit order at the higher price) but will receive the price improvement (of size τ) in case of execution.

Figure 2.4 around here

The third trader, observing the actions of the previous two as well as the state of the LOB $b_3 = [-1, 0]$, can undercut the second trader by placing a buy limit order at price p_3 . This order submission means the second trader will not receive execution of its buy limit order, as there is just the last trader left before the end of the trading day. The last trader will never opt for a limit order since no market orders may hit the LOB after him. Therefore the strategy set of the last trader contains only market orders or no trading.

²The model was solved with the following parameters: $\beta \sim \text{Unif}[-1, 1]$, $\tau = 0.05$, $v^H = 1.5$, $v^L = 0.5$, $\pi = 0.2$, $\alpha = 0$, $\epsilon = 10^{-4}$.

2.3.2 Multiple-Venue Market Equilibrium

In this section, I allow a fraction of traders α to have the access to DP. I show that the fact that the dark pool is available for some traders will alter the equilibrium trading strategies for all participants. The channel through which DP changes the equilibrium order submission probabilities is in the changes it brings to the limit order execution uncertainty and the ability not to reveal information about the pay-off.

Traders with no access to DP anticipate a shift in the order flow towards the dark market. This plays a key role in their willingness to submit market orders instead of limit orders. On the other hand, informed traders might choose to trade in the dark to reduce the information leakage to the market, and thus maintain their competitive advantage.

In the multiple venue market, both dark orders and IOC type orders become available for the fraction α of traders. The model is solved backwards, similarly to the procedure for the single-venue market. Additionally, I assume that, in the beginning, upon observing some hidden activity, traders assign equal probabilities to every possible state of DP. For example, if in the first period in the model there was no visible activity (limit order or market order) from the first trader, then the second trader's belief about the state of the dark pool is as follows:

$$DP_t = \begin{cases} 0 & \text{with probability } \frac{1}{3} \\ -1 & \text{with probability } \frac{1}{3} \\ 1 & \text{with probability } \frac{1}{3} \end{cases}, \quad (2.16)$$

where $DP_t = 0$ indicates the state where the first trader did not trade, $DP_t = -1$ the state where the first trader submitted the dark buy order, and $DP_t = 1$ the dark sell order. The submission of the IOC order will result in a market order since the dark pool opens empty.

After solving the model with the above assumption, I obtain the equilibrium order submission probabilities for every trader. As these order submission probabilities are common knowledge, some traders might find it optimal to choose a particular dark order to increase their expected utility. This will result in an update of the belief about the state of DP by all traders.

For example, in the state of the world where $V = v^H$, after the third trader does not display any visible activity (given that the first and the second trader's actions were unambiguous),

the last (informed) trader will update the initial belief in equation (2.16) as follows:

$$DP_4^* = \begin{cases} 0 & \text{with probability } \omega_{NT} \\ -1 & \text{with probability } \omega_{BD}, \\ 1 & \text{with probability } \omega_{SD} \end{cases} \quad (2.17)$$

where

$$\omega_i = \begin{bmatrix} \pi \\ 0 \\ 1 - \pi \end{bmatrix}^T \cdot \begin{bmatrix} \Pr(\varphi_i^{(3)}|v^H, \text{WA}) & \Pr(\varphi_i^{(3)}|v^H, \text{NA}) \\ \Pr(\varphi_i^{(3)}|v^L, \text{WA}) & \Pr(\varphi_i^{(3)}|v^L, \text{NA}) \\ \Pr(\varphi_i^{(3)}|E(V|\mathcal{F}_3), \text{WA}) & \Pr(\varphi_i^{(3)}|E(V|\mathcal{F}_3), \text{NA}) \end{bmatrix} \cdot \begin{bmatrix} \alpha \\ 1 - \alpha \end{bmatrix}, \quad (2.18)$$

and $i \in \{\text{NT}, \text{BD}, \text{SD}\}$. Here, the condition WA/NA indicated whether the third trader has or does not have an access to the dark pool. However, if the last trader is uninformed about the value of the final payoff, she will estimate the above probabilities as follows:

$$\omega_i = \begin{bmatrix} \pi \cdot \Pr(V = v^H|\mathcal{F}_3) \\ \pi \cdot \Pr(V = v^L|\mathcal{F}_3) \\ 1 - \pi \end{bmatrix}^T \cdot \begin{bmatrix} \Pr(\varphi_i^{(3)}|v^H, \text{WA}) & \Pr(\varphi_i^{(3)}|v^H, \text{NA}) \\ \Pr(\varphi_i^{(3)}|v^L, \text{WA}) & \Pr(\varphi_i^{(3)}|v^L, \text{NA}) \\ \Pr(\varphi_i^{(3)}|E(V|\mathcal{F}_3), \text{WA}) & \Pr(\varphi_i^{(3)}|E(V|\mathcal{F}_3), \text{NA}) \end{bmatrix} \cdot \begin{bmatrix} \alpha \\ 1 - \alpha \end{bmatrix} \quad (2.19)$$

Since the third trader can follow the belief updates of the last trader, she can then alter her order submission probabilities to get a higher expected utility. Eventually, there will be no further way the third trader can change her order submission probabilities to be better off. At this moment the belief of the last trader about the state of the dark pool is consistent with the strategies of the previous trader. I call this situation a local equilibrium.

Once I reach the local equilibrium, I go further backward and obtain strategies and the beliefs of the last two traders in the model, given that the second trader submitted a dark order. After local equilibriums are reached for each trader, similarly to the single-venue market, I re-examine the conditional probabilities that the final asset's pay-off is high ($V = v^H$). The procedure is repeated with the updated conditional probabilities until the model reaches the global equilibrium.

Figure 2.5 illustrates the algorithm to obtain the global equilibrium for the multiple-venue

market. The global equilibrium in the multiple venue market is reached when both beliefs about the DP state and conditional expectations of the final pay-off of the asset are consistent with the strategies of traders at every period.

Figure 2.5 around here

Figure 2.6 represents a part of the tree of the sequential game that the agents play during trading periods in the multiple-venue market³. Only actions of the last two traders are presented. In this example, after the first trader submits a buy limit order at price p_2 and the second trader submits a sell limit order at price p_3 , the state of LOB is $b_3 = [-1, 1]$ and $DP_3 = 0$.

Figure 2.6 around here

Observing \mathcal{F}_3 , trader 3 can deduce that the dark pool is still empty and might submit a dark order (for example, buy dark order) in case her private valuation of the asset is not high enough to justify a market order. If the trader doesn't have access to the dark pool and her private valuation is not high enough, the trader does not trade. In both cases, the state of the LOB is left intact but the state of the DP is different.

The last trader, however, cannot distinguish between the two states Ω_4 and Ω'_4 , and has to infer the state of the dark pool to assess the pay-offs of trading in the dark.

Figure 2.7 plots the equilibrium composition of the strategies the agents use during the trading periods. The figure depicts the probability with which traders choose market orders versus aggressive limit orders (those are placed to undercut the existing limit orders), non-aggressive limit, and dark orders. The dark orders category also includes the no trading option, which is by far dominating. Panel (a) presents the case for varying access to the DP while the proportion of informed traders stays constant. There is no noticeable variation in the average probabilities of choosing between market orders, limit orders, or dark ones as the dark pool becomes more accessible.

Figure 2.7 around here

In contrast, Panel (b) of Figure 2.7 shows changes in strategies composition when only the share of informed traders changes. With the increase in the share of the informed traders in the model,

³The model was solved with the following parameters: $\beta \sim \text{Unif}[-0.5, 0.5]$, $\tau = 0.05$, $v^H = 1.25$, $v^L = 0.75$, $\pi = 0.2$, $\alpha = 0.5$, $\epsilon = 10^{-4}$.

the fraction of aggressive (both market and limit) orders increases. When the unconditional probability to meet an informed trader surpasses 50%, the aggressiveness of orders reduces. The same shape can be observed for the fraction of dark orders. Even though, the no trading option is not better than any other dark order it is dominating the order submission probability.

2.4 Venue Competition

In this section, I discuss the competition between the limit order book and the dark pool. The competition happens in the order flow. In the single-venue market, every trader stays in the LOB while in the multiple-venue market traders can migrate to the DP.

To quantify the order flow to the dark pool, I follow Buti et al. (2017) and compute the Order Migration metric as the average across time periods probability that an order migrates to the DP:

$$OM = \frac{1}{T} \sum_{t=1}^T \mathbb{E}(\varphi_{BD} + \varphi_{SD} + \varphi_{BIOC} + \varphi_{SIOC} | \mathcal{F}_t), \quad (2.20)$$

where $T = 4$ is the number of time periods (and traders) in the model, $\mathbb{E}(\dots | \mathcal{F}_t)$ is the probability that a trader submits an order to the DP, given all available information \mathcal{F}_t .

Figure 2.8 illustrates how order migration changes with access to the dark pool. On Panel (a), we can see that the order migration monotonically increases with a more accessible dark venue. For $\pi = 0.6$, the order migration increases faster compared to lower levels of informed traders' presence. Interestingly, one might notice that for a moderate level of information asymmetry ($\pi = 0.3$), traders use dark orders less often than for low or high levels.

This pattern is better visualized in Panel (b) of Figure 2.8. For a broad range of DP availability, we observe the initial decline in orders' migration to the dark. As the information asymmetry grows further, traders come back to the dark pool, and the order migration increases. This creates a “smile” pattern of order migration versus information asymmetry.

Figure 2.8 around here

It is insightful to know which traders send orders to the DP more often. Both traders can benefit from using dark orders as there is no fee for submitting an order to the DP. Uninformed traders can get the price improvement by executing at the current midquote. In addition to

this, informed traders can maintain the informational advantage by hiding their activity in the dark pool.

On the other hand, as pointed out by Zhu (2014), informed orders may cluster on one side of the market. Therefore, the non-execution risks for these orders increase. The trade-off between price improvement, information concealing, and low execution risks determines the degree to which informed traders utilize the dark pool.

For this model, I calculate the probability that the order was submitted by an uninformed trader, given that it was sent to the dark pool. Figure 2.9 shows the relationship between this conditional probability and the share of the informed traders in the model. The dashed line represents the situation when there is no crowding-out effect. When 20% of traders possess insider information and there is no discrimination in the access to the dark pool, one would expect to observe 20% of dark orders being sent by informed investors and 80% by the uninformed.

Figure 2.9 around here

As the blue line shows, there is a crowding-out effect in this model. Uninformed traders send their orders to the DP much rarely in the presence of information asymmetry. There is a steep decrease in the conditional probability of an uninformed trader submitting a dark order at low levels of π . It signals that the majority of the uninformed market participants take the risk of trading with the informed in the DP very seriously. As a result, with about 90% of uninformed traders in the model, we observe only 40% of all dark orders come from uninformed.

At higher levels of information asymmetry, there is a linear, less steep decline in the conditional probability that the dark order was submitted by an uninformed trader. In this region, uninformed traders face the trade-off between joining a long queue and trading with an informed participant.

2.5 Empirical Implications

2.5.1 Price Discovery

The first set of empirical implications involves price discovery. Overall, price discovery is known to be a process of determining the proper price of an asset. This process may depend on a

variety of factors like supply and demand dynamics, risk attitudes of market participants, and general economic environment.

An important factor in the price discovery process is the information flow. Market participants with more precise, high-quality information have an advantage as they can act before this information reaches others. By their trading, informed participants reveal part of their private information to the market. By observing the order flow, uninformed traders update their expectations about the future value of the asset.

In this model, only the information incorporation factor is considered when describing the effectiveness of the price discovery process. Before the trading starts, uninformed traders are aware of the unconditional volatility of the value of the final pay-off. After each trade or limit order submission, uninformed participants can compute the volatility of the pay-off conditional on the actions of previous traders. As $\sigma_{V|\mathcal{F}_t} \leq \sigma_V, \forall t$, the order flow incorporates private information into prices.

I measure price discovery by a percentage decrease in conditional volatility relative to the unconditional volatility (which is also the fundamental asset volatility) from the viewpoint of the last trader in the model. Thus, price discovery of -50% means that the conditional volatility of the final pay-off for the last trader is two times less than its fundamental volatility.

Price discovery for different degrees of dark pool accessibility is presented in Figure 2.10. Panel (a) displays the average price discovery. We can observe that with a more accessible dark pool, the conditional volatility of the final pay-off decreases less. This fact suggests a less efficient price discovery process with easily available DP. However, in relative terms, the average price discovery didn't change much when the α parameter changes from zero to one. This is an indication of a largely limited impact of the DP on the limit order markets in their ability to determine the appropriate price.

Figure 2.10 around here

I also present the results for the best possible price discovery on panel (b) of Figure 2.10. For all cases, the sequence of three consecutive market orders reveals most of the information and delivers the best price discovery. Surprisingly, we may observe an increase in price discovery for more easily accessible DP. In relative terms, this result is almost negligible.

2.5.2 Market Quality

In this section, I report the results of the three most common measures of execution quality employed in market microstructure literature. The quality of execution of market participants' orders is crucial for financial regulation. In financial markets, execution quality is closely linked to market liquidity. According to Foucault et al. (2013b), liquidity is the degree to which an order can be executed within a short time at a consensus price. This price is, in most cases, tied to the middle of the National Best Bid and Offer (NBBO). If the execution price is closer to the midpoint of the NBBO and there are enough shares offered so that the market order doesn't push the price, the security is considered more liquid from the point of view of the aggressive participant (liquidity demander).

On the other hand, sufficient depth of the market (number of shares offered at different price levels) may cause long waiting times for those who decide to submit a limit order. This would make the liquidity providers worse off as they have to wait until their offer is taken. That is why taking into account different measures is crucial to obtain a broad view of execution quality.

Spread

Bid-ask spread is the most common measure of execution quality. The quoted bid-ask spread is the difference between the best ask and the best bid price, $\text{spread}_t = p_t^{\text{ask}} - p_t^{\text{bid}}$. For this model, I calculate the average across periods quoted bid-ask spread as follows:

$$\text{avg. spread} = \frac{1}{T-1} \sum_{t=1}^{T-1} \sum_{\varphi \in \Phi} \text{spread}_{t,\varphi} \cdot P(\varphi | \mathcal{F}_t), \quad (2.21)$$

where $T = 4$ is the number of time periods in the model, $\Phi = \{\varphi_{BM}, \varphi_{SM}, \dots, \varphi_{NT}\}$ is a set of all strategies, $\text{spread}_{t,\varphi}$ is the quoted bid-ask spread after implementing strategy φ by trader t , and $P(\varphi | \mathcal{F}_t)$ is the probability that trader t uses this strategy φ given all available information \mathcal{F}_t .

In equation 2.21, I do not account for the last (fourth) trader because this trader only submits market buy, sell, or an aggressive dark order. Therefore, the last trader is not capable of providing liquidity in any market conditions and does not bring meaningful variation in the average spread measure.

Figure 2.11 around here

Figure 2.11 reports changes in the average spread in the model when (a) access to the dark pool changes and (b) the share of informed traders changes. The figure shows that there is no change in the average bid-ask spread when varying the access to the DP. This may be explained by the equivalent migration of market orders to the dark as well as limit orders. In this case, the dark venue does not impair the execution quality on the exchange.

On the other hand, panel (b) of Figure 2.11 shows that, at first, the bid-ask spread increases with the increase in the share of informed traders. It reaches its maximum at $\pi \approx 0.5$ and then starts to decline with approximately the same but negative slope. These changes in the bid-ask spread are aligned with the process of Order Migration, in Figure 2.8, panel (b). As informed traders use DP relatively more often, in the times of low order migration, their aggressive (market) orders widen the spread on the exchange. When the order migration increases again ($\pi > 0.5$), the market orders' pressure leaves the exchange and lets the spread recover.

Market Depth

A deep market is one where large market orders do not have a greater price impact than small orders. The depth of the market is, therefore, measured by the price impact of trades. As in this model, traders can only submit orders for one share, there is not much sense in calculating price impacts. Instead, I compute the depth by directly assessing the average number of shares outstanding in the LOB at every period. Thus, the measure of the market depth is the following:

$$\text{avg. depth} = \frac{1}{T-1} \sum_{t=1}^{T-1} \sum_{\varphi \in \Phi} \text{depth}_{t,\varphi} \cdot P(\varphi|\mathcal{F}_t), \quad (2.22)$$

where $T = 4$ is the number of time periods in the model, $\Phi = \{\varphi_{BM}, \varphi_{SM}, \dots, \varphi_{NT}\}$ is a set of all strategies, $\text{depth}_{t,\varphi}$ is the combined amount of shares outstanding on buy and sell sides of the market after implementing strategy φ by trader t , and $P(\varphi|\mathcal{F}_t)$ is the probability that trader t uses this strategy φ given all available information \mathcal{F}_t .

Figure 2.12 reports the average across periods market depth when (a) varying the access to the dark pool, and (b) varying the share of informed traders.

Figure 2.12 around here

The figure shows that there is a strong correlation between the quoted bid-ask spread measure and the market depth measure. Similar to the bid-ask spread, the average depth does not change when DP becomes more easily available to traders. However, when the fraction of informed traders increases, depth declines at first, and then, after the share of informed traders surpasses 50%, recovers.

Fill Rate

Both bid-ask spread and market depth measures proxy for the execution quality of market orders. Contrary to this, the analysis of the execution quality of limit orders is often skipped. Dugast (2020) studies the relationship between the execution delay of limit orders and market depth. He finds that despite both measures co-varies negatively with welfare, the execution quality of limit orders may dominate in the welfare outcome. The conclusion is that the market depth is inefficient, and cannot be the only measure to infer investor's welfare variations. The execution delay is crucial for the comprehensive assessment of the market quality.

To account for the execution quality of limit orders, I estimate the average across periods fill rate of such orders. The fill rate is an alternative measure to the execution delay of limit orders. Close to one value of the fill rate would imply a short waiting time for a limit order while close to zero value of the fill rate means long waiting in the LOB.

For each but the last trader in the model, I compute the fill rate value as the probability that a liquidity-providing order would receive execution in the next trading periods. I then average the fill rates across those traders to obtain a general view of the fill rates in the model:

$$\text{avg. FR} = \frac{1}{T-1} \sum_{t=1}^{T-1} \left(\frac{\sum_{\varphi \in \Phi_L} \mathbb{E}(\varphi \cdot \mathbb{I}_{\text{exec}} | \mathcal{F}_t)}{\sum_{\varphi \in \Phi_L} \mathbb{E}(\varphi | \mathcal{F}_t)} \right), \quad (2.23)$$

where $T = 4$ is the number of periods in the model, $\Phi_L = \{\varphi_{BL}, \varphi_{SL}, \varphi_{BD}, \varphi_{SD}\}$ is a set of liquidity providing strategies and \mathbb{I}_{exec} is the indicator variable of the execution of order φ .

Figure 2.13 reports the average across traders fill rate when (a) varying the access to the dark pool, and (b) varying the share of informed traders. Panel (a) suggests that the fill rate increases monotonically when more traders have access to the dark pool. This is possible due to the smart routing system of the IOC orders. Impatient traders can guarantee the execution of their aggressive orders even when submitting those to the dark pool. Part of those orders

will still be routed to the LOB and improve the execution quality of limit orders.

Figure 2.13 around here

Panel (b) shows that, at first, an increased share of informed traders improves the execution quality by intensifying market orders' flow. There is a point, however, where a further increase in the share of informed traders will create a noticeable asymmetry in the order flow, reducing the execution rate on one side of the market.

2.5.3 Welfare

Since all participants in the model have private valuations, a trade between two parties generates welfare gains. Further, I study the impact of the DP on the realization of welfare gains of traders.

Following the work of Goettler et al. (2005), Degryse et al. (2009), and Buti et al. (2017), I compute the welfare of a trader of type $j \in \{v^H, v^L, E(V|\mathcal{F}_t)\}$, and with DP access $k \in \{\text{NA}, \text{WA}\}$ as follows:

$$W_{t|j,k} = \int_{-1}^1 \mathcal{U}_t^*(\beta) d\beta, \quad (2.24)$$

where $\mathcal{U}_t^*(\beta)$ is the utility of the optimal strategy at time t , obtained by solving the maximization problem (2.11).

The total welfare aggregated for all traders is then calculated as:

$$W = \sum_{t=1}^T \left(\begin{bmatrix} \pi/2 \\ \pi/2 \\ 1 - \pi \end{bmatrix} \right)^T \cdot \begin{bmatrix} W_{t|v^H, \text{WA}} & W_{t|v^H, \text{NA}} \\ W_{t|v^L, \text{WA}} & W_{t|v^L, \text{NA}} \\ W_{t|E(V|\mathcal{F}_t), \text{WA}} & W_{t|E(V|\mathcal{F}_t), \text{NA}} \end{bmatrix} \cdot \begin{bmatrix} \alpha \\ 1 - \alpha \end{bmatrix} \right). \quad (2.25)$$

Figure 2.14 illustrates changes to the aggregate welfare in the model when (a) varying the access to the dark pool, and (b) varying the share of informed traders. Panel (a) depicts the fact that the total welfare of market participants is highly stable for all levels of dark pool availability. We observe a slight increase in welfare, driven by the increase in order migration.

Figure 2.14 around here

Panel (b) of Figure 2.14 shows a more sizeable increase in the aggregate welfare as the fraction of informed traders in the model grows. A greater number of informed traders would imply more market orders and, thus, faster realizations of the gains from trade.

As a matter of fact, Figure 2.15 shows that both informed and uninformed traders experience a relative reduction in the gains from trade as the fraction of informed increases. This welfare reduction is in line with an increase in the average fill rate, bid-ask spread, and a decrease in the market depth. However, as the informed receive higher gains from trade, the overall welfare in the model increases monotonically with π .

Figure 2.15 around here

2.6 Conclusion

This paper analyzes the impact of dark pool operating alongside a transparent limit order exchange on market quality and price discovery. In the four-period model with asymmetric information about the final pay-off of the risky asset, fully rational, risk-neutral traders can choose order type and venue to send their orders to. I find that the order flow migrates to the dark pool in the last periods, as the limit order book fills.

When orders are directed to the opaque venue, the process of price discovery slows down. However, in relative terms, the effect is close to negligible and may vary in nature (benefiting price discovery) depending on the immediacy demands of investors. As limit orders provide a substitute for dark orders, venue competition plays a significant role in the impact of dark trading on price discovery and market quality. When information asymmetry increases the order migration forms a U-shape and the uninformed investors are quickly crowded out from the dark pool.

The model suggests that a dark pool affects market quality mildly. The execution quality of market orders measured by the average quoted bid-ask spread and the average depth is stable in the wide range of the dark pool availability. Also, the quality of limit orders, measured by the average fill rate, slightly increases with the more available dark venue. Overall, the total welfare remains leveled for various degrees of dark pool availability.

The developed model is flexible and allows for a wide range of analyses that interest policy-makers and regulators. However, future work can extend the model by addressing the following

caveats. Firstly, since traders enter the model only once, I do not allow for price manipulation. If a trader can submit an order several times during the trading day, she might choose to trade first on the lit exchange to move the market and then execute later in the dark pool at a more comfortable price.

Secondly, one might be interested in endogenizing private information acquisition and dark pool participation. By introducing a cost of obtaining perfect information about the final payoff of the asset and a cost of accessing the dark pool it will be possible to identify an equilibrium level of dark trading.

Finally, more types of orders can be added to the model to bring it closer to the current state of financial markets where the competition between transparent and dark venues is more subtle. Some exchanges allow traders to submit undisplayed orders that offer a price improvement and execute against the visible order flow. The introduction of such order types to the present model would be interesting future research.

References

- Bayona, Anna, Ariadna Dumitrescu, and Carolina Manzano (2020). “Information and Optimal Trading Strategies with Dark Pools”. *Working Paper*.
- Brolley, Michael (2020). “Price Improvement and Execution Risk in Lit and Dark Markets”. *Management Science* 66.2, pp. 863–886.
- Buti, Sabrina, Barbara Rindi, and Ingrid M. Werner (2017). “Dark pool trading strategies, market quality and welfare”. *Journal of Financial Economics* 124.2, pp. 244–265.
- Degryse, Hans, Mark Van Achter, and Gunther Wuyts (2009). “Dynamic order submission strategies with competition between a dealer market and a crossing network”. *Journal of Financial Economics* 91.3, pp. 319–338.
- Dugast, Jérôme (2020). “Inefficient Market Depth”. *Working Paper*.
- Foucault, Thierry, Marco Pagano, Ailsa Roell, and Ailsa Röell (2013b). *Market liquidity: theory, evidence, and policy*. Oxford University Press.
- Goettler, Ronald L, Christine A Parlour, and Uday Rajan (2005). “Equilibrium in a dynamic limit order market”. *Journal of Finance* 60.5, pp. 2149–2192.

- Kyle, Albert S (1985). “Continuous auctions and insider trading”. *Econometrica*, pp. 1315–1335.
- Ye, Mao (2012). “Price manipulation, price discovery and transaction costs in the crossing network”. *Working Paper*.
- Zhu, Haoxiang (2014). “Do dark pools harm price discovery?” *Review of Financial Studies* 27.3, pp. 747–789.

Tables and Figures

Table 2.1. Available order types.

This table presents the orders available to traders for submission to the limit order book (LOB) and to the dark pool (DP). Only fraction α of traders has access to the dark pool. Market sell/buy order executes immediately at the best available bid/ask price. Limit order should specify the price, at which the trader wants to buy/sell the shares. The price is defined on a grid $p \in [p_1, p_2, p_{3,4}]$. Levels 1 and 4 of the grid are occupied by the crowd that immediately absorbs any amount of the open interest. Dark orders execute at the midprice that is determined by the available liquidity in the LOB. They require the counterparty for the trade to happen, otherwise, the order will wait for the execution in the DP according to the time priority rule. Dark Immediate or Cancel (IOC) orders are similar to the usual dark orders, but the execution of such orders is guaranteed by the routing technology. If the dark IOC order cannot be executed in the dark immediately at the midprice, it will be routed to the LOB and executed there against the best offer available.

Description	Notation
<i>Orders sent to the LOB</i>	
Market order (buy or sell)	$\varphi_{BM}, \varphi_{SM}$
Limit order (buy or sell)	$\varphi_{BL}(p_j), \varphi_{SL}(p_j), \quad j = 2, 3$
No trade	φ_{NT}
<i>Orders sent to the DP</i>	
Dark order (buy or sell)	$\varphi_{BD}(\tilde{p}_{Mid,t}), \varphi_{SD}(\tilde{p}_{Mid,t})$
Dark IOC order (followed by market order if cancelled)	$\varphi_{BIOC}(p_{Mid,t}), \varphi_{SIOC}(p_{Mid,t})$

Figure 2.1. Dark trading in USA and Europe (Source: Rosenblatt Securities).

The Figure plots dynamics in the dark pools' trading volume share in Europe and the United States for the period from February 2020 to March 2021. A rather stable trend around 13% and 16% of the total volume traded in the dark can be observed for the US and Europe, respectively.

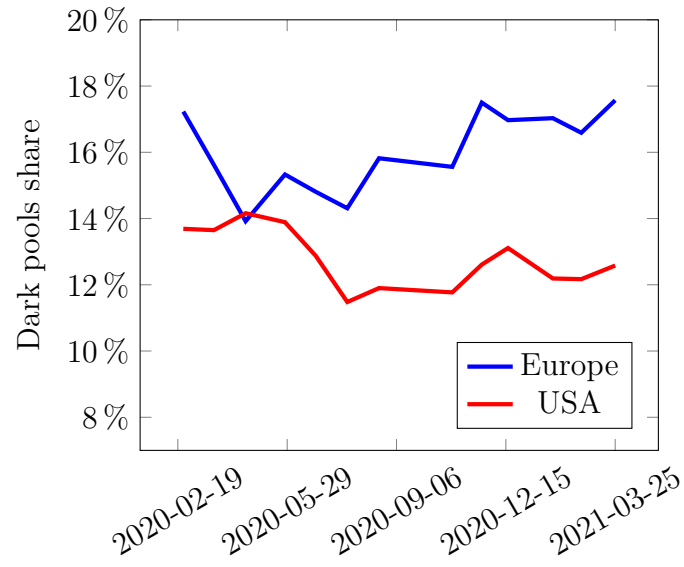


Figure 2.2. Limit orders in the limit order book (LOB).

The graphs below, show the possible state of the LOB. The horizontal axis represents the price levels that are available for order submission. The difference between the two neighboring price levels constitutes the minimum increment of the price (tick size). The vertical axis represents the amount of shares available at each price level. Negative (positive) quantity reflects available limit orders to buy (sell). Panel A, shows the situation when all levels of the LOB are occupied with limit orders so the next trader can either join the line of limit orders or choose to submit a market order. On Panel B, there are no orders at price level p_2 , and only buy limit orders can take this place in the next periods as “crossed market” is not allowed.

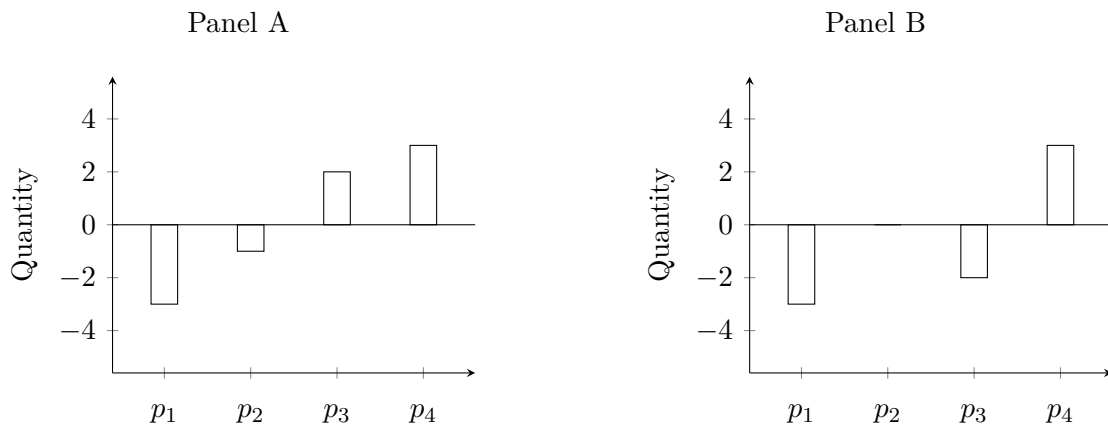


Figure 2.3. Procedure to obtain an equilibrium in a single-venue market.

The Figure shows the steps taken to solve numerically the model of the Limit Order Book (LOB) trading when the dark pool is not available. At the first step, each trader assumes that the value of the security is high with some probability. Then the strategies of all traders are calculated by backward induction. If the strategies of all traders are not aligned with their initial beliefs (up to a certain small $\epsilon > 0$), those beliefs are updated and the procedure is repeated.

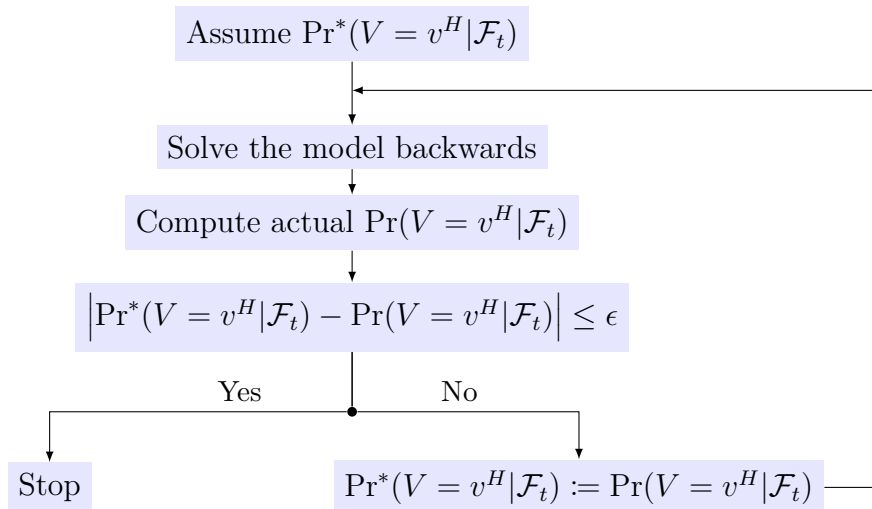


Figure 2.4. The extensive form of the trading game in the single-venue market.

When only the limit order book (LOB) is available, the traders may choose between market orders, limit orders, or no trading option. Thus, every trader can observe the actions of all previous traders and update her beliefs. One possible realization of the strategic game is marked by blue lines. The variable b_i shows the state of the LOB at time period i where the first element shows the number of shares available at price level p_2 and the second element - at price level p_3 .

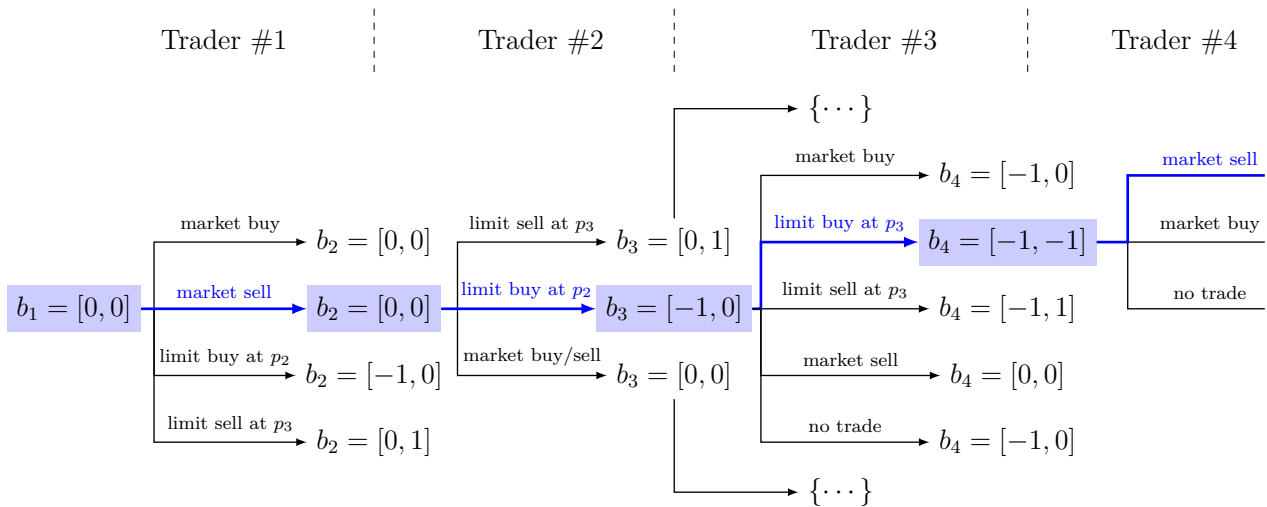


Figure 2.5. Procedure to obtain an equilibrium in a multiple-venue market.

The Figure shows the steps taken to solve numerically the model of the Limit Order Book (LOB) trading when the dark pool (DP) is available for α share of the traders. As in the single-venue market (see fig 2.3), the procedure starts with an assumption about all traders' beliefs of the true value of the asset. Next, the model is solved backwards in the following steps. Each trader aligns her strategies with the fact that the next ones can have an access to the DP with probability $\alpha > 0$. When the beliefs that the next traders will opt for the DP are within a small distance $\epsilon_1 > 0$ of the actual probability of the dark trading happening, the previous trader in line goes through the same calibration exercise. After the beliefs of all traders are in line, the actual probability for the value of the asset being high is computed and adjusted if too far from the initial assumption. The calibration continues until the equilibrium is reached in both beliefs about the (i) true value of the asset and (ii) the probability of dark trading.

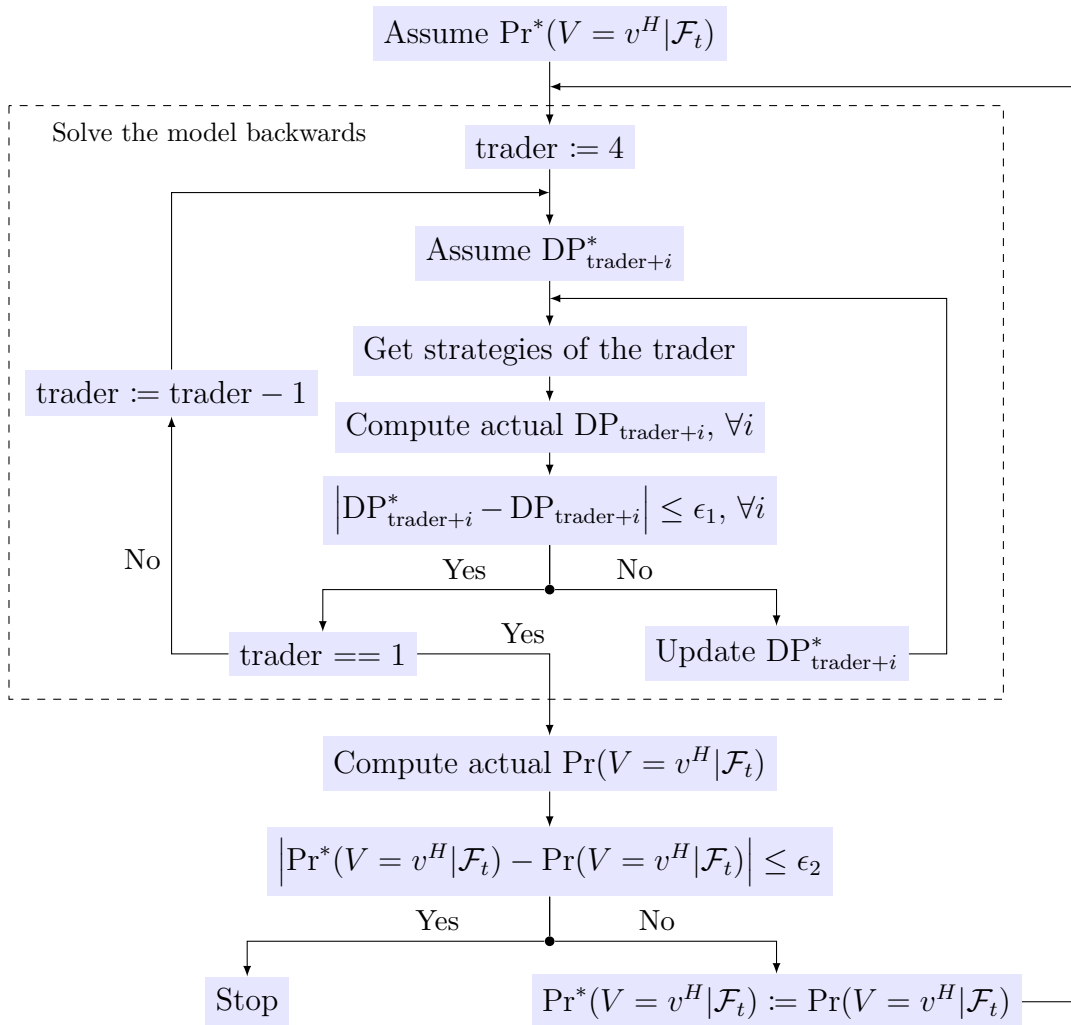


Figure 2.6. The extensive form of the trading game in the multiple-venue market.

When both a limit order book (LOB) and a dark pool (DP) are available, traders cannot distinguish dark orders and no trading option with certainty. The variable Ω_t shows both the state of the LOB and the true state of the DP that is not directly observable by any trader. The Figure below illustrates how Trader 3 strategically decides to submit dark buy order, knowing that the next Trader 4 cannot observe her actions and will be able to trade with her only if both of them have access to the DP. In this case, Trader 3 faces a trade-off between the uncertainty of the execution and the more favorable execution price.

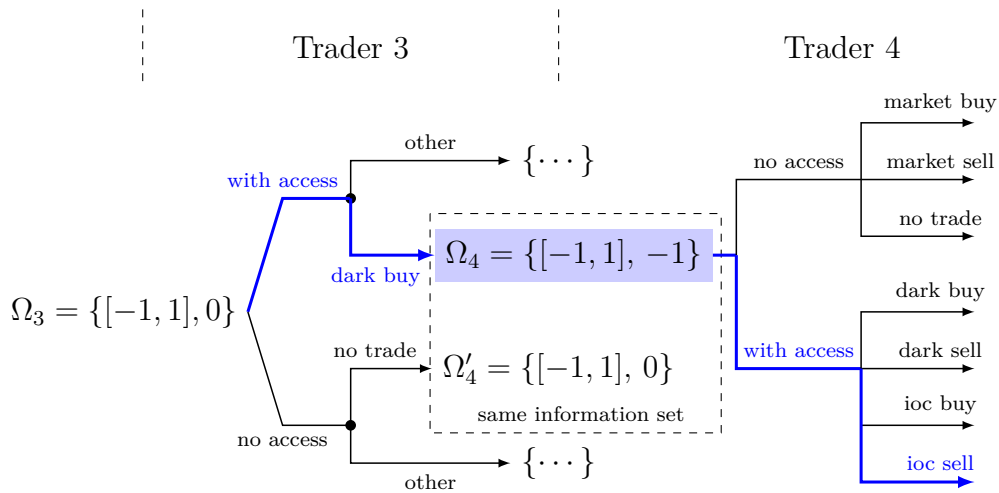
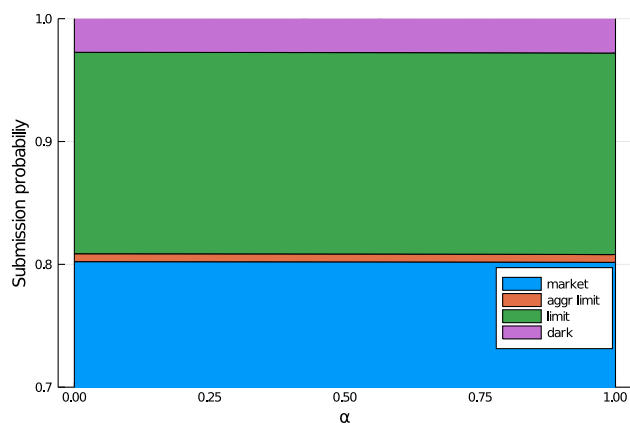
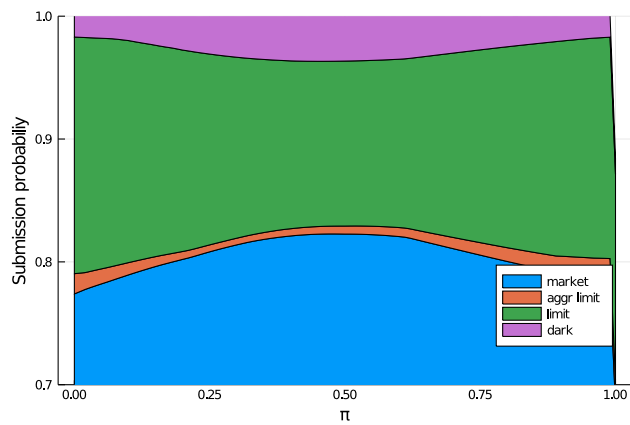


Figure 2.7. Equilibrium order submission strategies composition.

This Figure shows the probability of submitting different types of orders based on (a) availability of the dark pool (DP), and (b) the share of informed traders among the market participants, π . Panel (a) suggests that the strategies of traders do not change that much if the DP becomes more widely available while the share of informed traders stays constant. On the other hand, panel (b) shows that trading via dark orders and market orders first increases along with the increase in the number of informed traders but as the share of those surpasses 50%, this effect reverts.



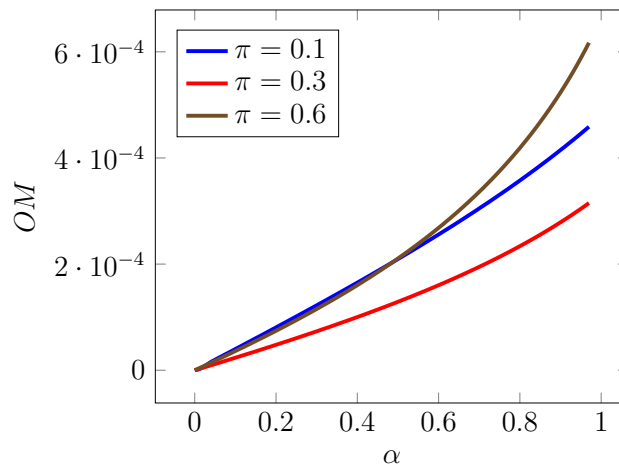
(a) Varying α , ($\pi = 0.2$)



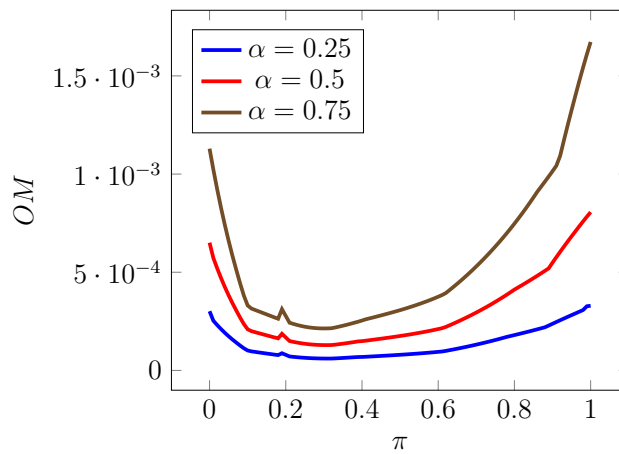
(b) Varying π , ($\alpha = 0.5$)

Figure 2.8. Order Migration for (a) various dark pool (DP) availability α , and (b) information asymmetry π .

Order Migration, defined by equation 2.20, is, essentially, an average across periods probability that a dark order will be submitted. Naturally, if the DP becomes more available for trading, the level of order migration goes up. It, however, does not reach high values as liquidity provision remains more beneficial during most of the trading day. Pane (b) shows how the order migration changes when the share of informed traders increases. For all levels of the DP availability, one can notice a ‘smile’ shaped curve of the order migration. The initial decrease in the order migration may be associated with a risk of matching with the informed trader who is trying to conceal her informational advantage. The subsequent increase happens when π increases over 0.5 as the liquidity builds up in the LOB.



(a) $\pi \in \{0.1, 0.3, 0.6\}$



(b) $\alpha \in \{0.25, 0.5, 0.75\}$

Figure 2.9. The probability that the dark order came from an uninformed trader.

The Figure shows the conditional probability that the order was submitted by the uninformed trader given that the order was sent to the dark pool (DP). As the graph lies below the -45° line, one can deduce that uninformed traders are hesitant to use the DP even if they have an access to it. With the increase of the share of informed traders in the model from 0 to around 0.15, a considerable crowding-out effect of the uninformed traders from the DP is observed. As the uninformed ones are the majority, this effect can explain the initial decrease in the order migration. With a further increase in the share of the informed traders, the conditional probability decreases only linearly and reaches zero when all the agents are informed about the true value of the asset.

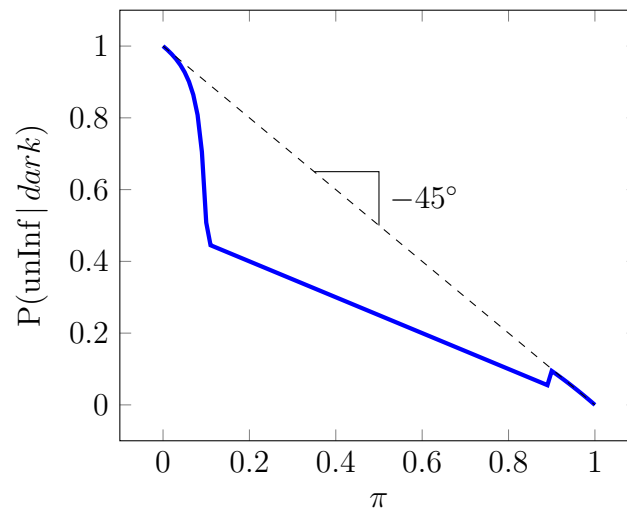
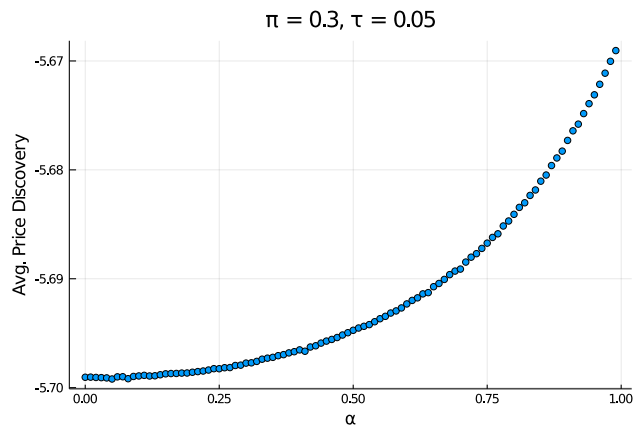
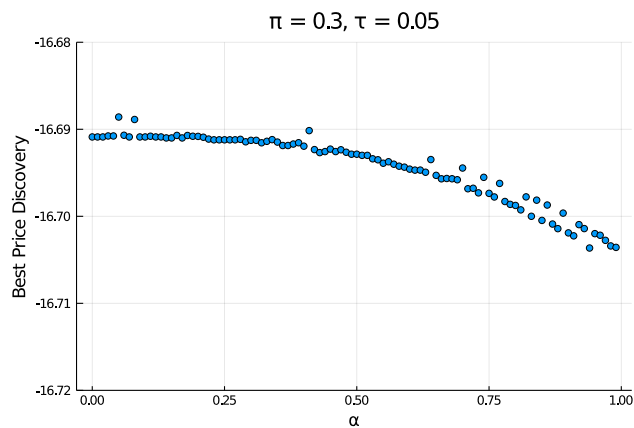


Figure 2.10. Price discovery for different levels of α .

The Figure below shows the dynamics in price discovery measure when the dark pool (DP) becomes more available. Panel (a) depicts the average price discovery defined as an average percentage change in conditional volatility of the asset's price, relative to the unconditional volatility from the point of view of the last trader in the model. The panel shows a less efficient price discovery mechanism for the markets with easily available DPs. Panel (b) shows the best possible price discovery dynamics. As the market order reveals the biggest amount of information, three consecutive market orders will provide the best price discovery to the last trader. Unexpectedly, one can observe, an increase in price discovery for more accessible DP. In absolute terms, the result is, however, negligible.



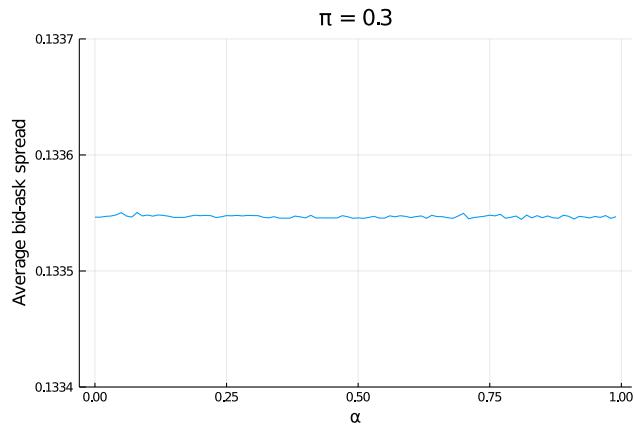
(a) The average price discovery



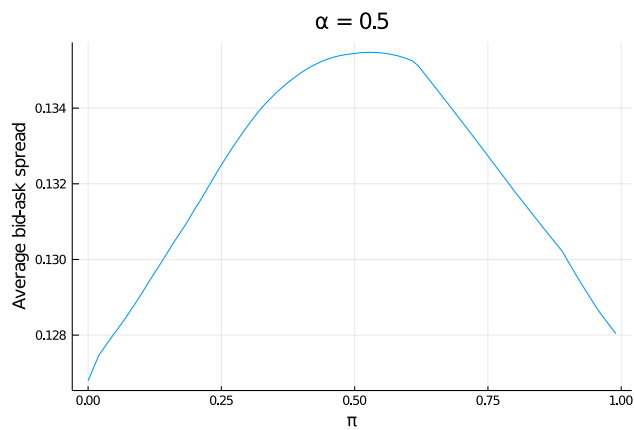
(b) The best price discovery

Figure 2.11. Average quoted bid-ask spread.

The figure below presents the dynamics in the average quoted bid-ask spread when the dark pool (DP) becomes more available (panel (a)), and when the fraction of informed traders in the model increases (panel (b)). The spread is defined by the difference between the best bid and ask prices. Panel (a) suggests that there is no effect of the availability of the DP on the average spread. This may be explained by the equivalent migration of market orders and limit orders to the DP. Panel (b) suggests a hump-shaped change in the average spread as the probability of informed trading increases. This effect is aligned with the dynamics of order migration, as informed traders use the DP more often.



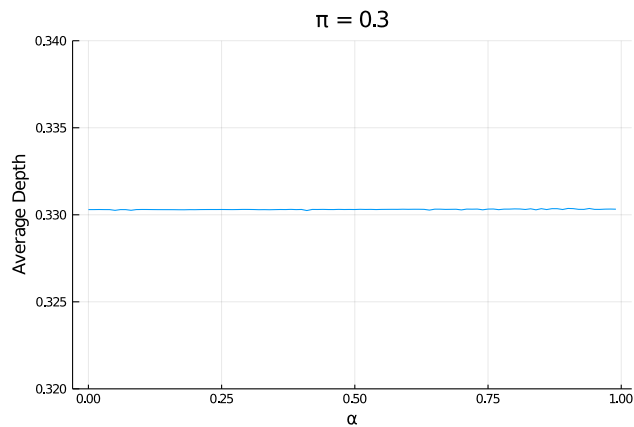
(a) Varying α , $\pi = 0.3$



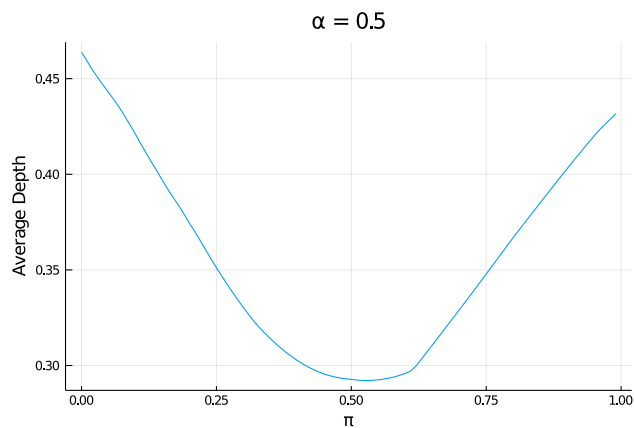
(b) Varying π , $\alpha = 0.5$

Figure 2.12. Average market depth.

The Figure below shows the dynamics in the average market depth when the dark pool (DP) becomes more available (panel (a)), and when the fraction of informed traders in the model increases (panel (b)). The depth was computed as the average number of shares outstanding in the limit order book (LOB) at every period. The figure shows that there is a strong correlation between the quoted bid-ask spread measure and the market depth measure. Similar to the bid-ask spread, the average depth does not change when DP becomes more easily available to traders. However, when the fraction of informed traders increases, depth declines at first, and then, after the share of informed traders surpasses 50%, recovers.



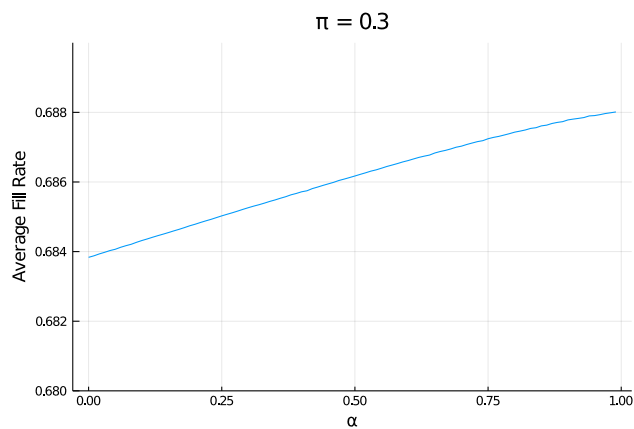
(a) Varying α , $\pi = 0.3$



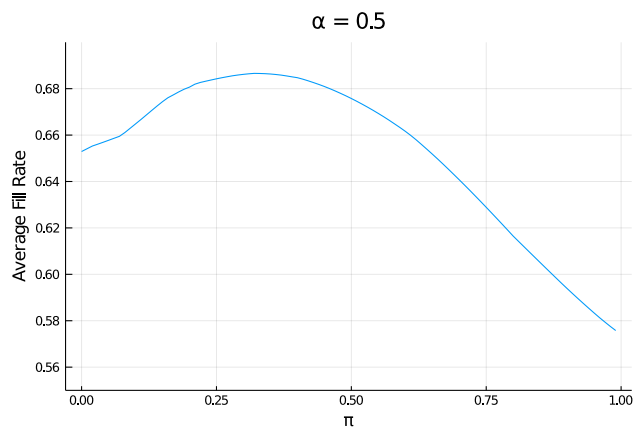
(b) Varying π , $\alpha = 0.5$

Figure 2.13. Average fill rate.

The Figure below shows the dynamics in the average fill rate when the dark pool (DP) becomes more available (panel (a)), and when the fraction of informed traders in the model increases (panel (b)). The fill rate is the probability that a liquidity-providing order would receive an execution in the next trading periods, before the end of the trading day. Panel (a) suggests that the fill rate increases monotonically when more traders have access to the DP. Panel (b) shows that, at first, an increased share of informed traders is associated with an improvement in the execution quality (possibly by intensifying market orders' flow). A further increase in π will be associated with a reduced execution rate on one side of the market.



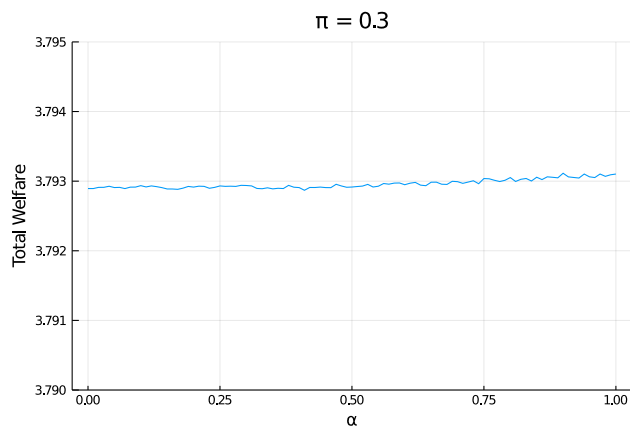
(a) Varying α , $\pi = 0.3$



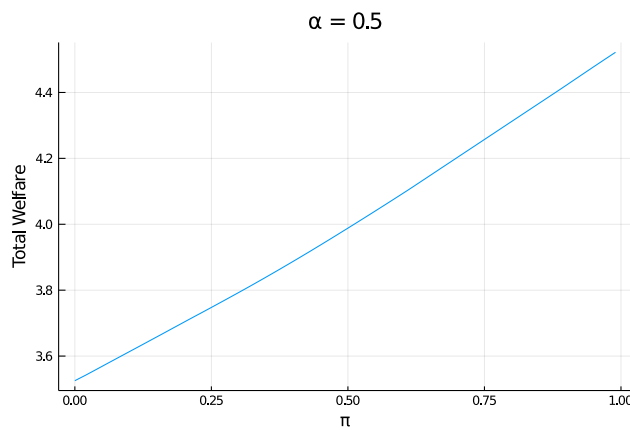
(b) Varying π , $\alpha = 0.5$

Figure 2.14. Aggregate welfare.

The Figure below shows the dynamics in the aggregate welfare when the dark pool (DP) becomes more available (panel (a)), and when the fraction of informed traders in the model increases (panel (b)). Panel (a) suggests that the total welfare of market participants is highly stable for all levels of dark pool availability. One can observe a slight increase in welfare mainly driven by the increase in order migration. Panel (b) shows a more sizeable increase in the aggregate welfare as the fraction of informed traders in the model grows. A greater number of informed traders would imply more market orders, and, thus, faster realizations of the gains from trade.



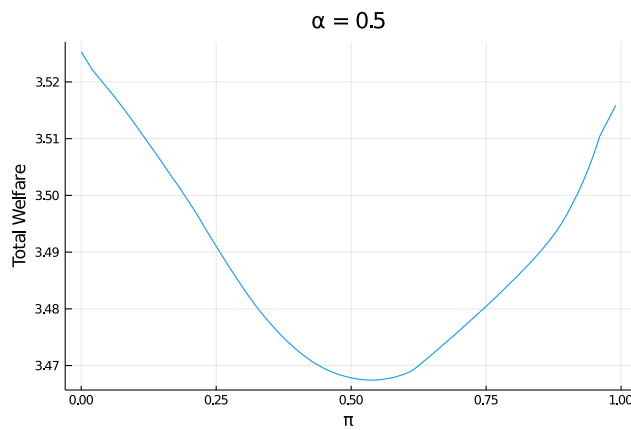
(a) Varying α , $\pi = 0.3$



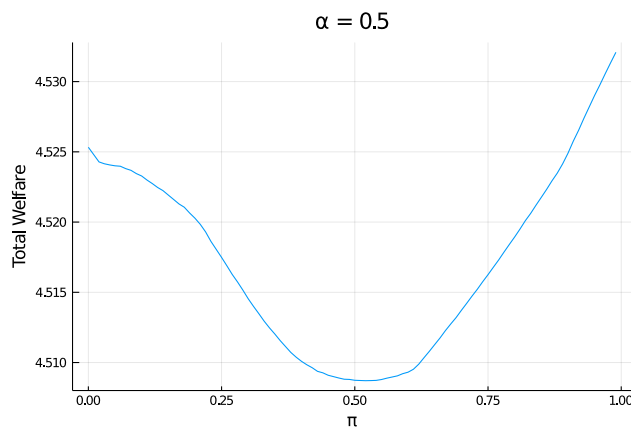
(b) Varying π , $\alpha = 0.5$

Figure 2.15. Aggregate welfare of traders with different information sets.

The Figure below shows the dynamics in the aggregate welfare of informed and uninformed traders when the fraction of informed traders in the model increases. Both panels, for informed and uninformed traders, suggest that all traders experience a relative reduction in the gains from trade as the fraction of the informed ones increases. The uninformed traders suffer because they become an easy target of the informed ones, while informed traders are worse off because their competitive advantage becomes less unique. However, as the informed traders receive higher gains from trade, the overall welfare in the model increases monotonically with π .



(a) Uninformed traders



(b) Informed traders

Chapter 3

Machine Learning and Market Microstructure Predictability

3.1 Introduction

Digitalization, which happened in the last decades, generated enormous piles of data that grow faster every day. The impact of big data is starting to be noticed in academic research in economics and finance. With the possible benefits of the abundance of data, emerge many questions, that are related to computer science and engineering. What does big data in finance really mean, and how can economists benefit from it?

In the finance literature, “big data” is usually associated with the following properties. First, the data should be large in an absolute or relative sense to capture relevant economic activities that are not present in the small dataset. Second, “big data” is often high-dimensional. It has many variables relative to the sample size. Finally, the data has a complex structure relative to the classical row-column format. It is the intricate structure of the data that allows for capturing economic activities that can not be measured with traditional data.

Natural examples of big data span from pictures, and text to audio and video files. A particular example, that is related to the current research, is an order-message level market microstructure data. This data is usually large as the exchange receives millions of trade and order submission messages every day, it is also high dimensional and complex as it is a source for limit order book reconstruction. To use efficiently the order message data, one requires to employ the “toolbox”, designed to account for the complexity of big data.

Machine learning is currently a common solution to the challenges imposed by the complex data, and it is used in economic and finance research more frequently. Machine learning techniques are quite diverse but, in general, they work well when (i) the approach towards the economic questions implies using a lot of variables, (ii) the variables are interconnected in a highly non-linear way, and (iii) prediction is more valuable than the inference.

In this paper, I further stress the importance of machine learning techniques for market microstructure research. Initially, it was not obvious that these methods can bring an improvement to the researchers' ability to predict how interactions between market participants affect price and liquidity dynamics. Economists believed that simple approaches can perform reasonably well, but it is becoming more obvious that the ability of machine learning to treat complex data using nonparametric algorithms can identify patterns that the traditional models can never capture.

For the purpose of this study, I take several machine learning models like feed-forward neural networks, long short term memory networks, gradient boosted machines, and fit them using the high-frequency data from the Nasdaq exchange. The data allows for calculation of various microstructure features that serve as predictors for the set of target variables that describe the market in the following aspects.

First of all, I try to predict short-term returns. Returns prediction is a cornerstone of the finance literature, that can also be an indication of market efficiency. Secondly, as a proxy of the general market stability, I predict the occurrence of mini flash crashes. These events, also called as "extreme price movements" or "price jumps", emerged with the digitalization of trading and the takeover of the markets by computer algorithms. Next, I check the performance of the models' forecasts for the quoted spread as the most widely used measure of market liquidity. Finally, I try to predict the short period market volatility by targeting the realized volatility measure.

In doing this, I leverage on the previous works¹ that applied machine learning methods for the purpose of predictions of microstructure variables. My addition to this previous work resides in the fact that I not only test the forecasting performance based on a way of data sampling, but also compare the accuracy of the models during normal times and crisis times as the data covers the period of the COVID-19 outbreak. Moreover, contrary to Easley et al.

¹For example, Mullainathan and Spiess (2017) and Easley et al. (2021).

(2021), I pursue, to my view, a more challenging goal of a regression forecast rather than a classification exercise.

To analyze the performance of machine learning methods, I use the out-of-sample R^2 measure for the models' output, and, to statistically compare models between each other, I employ the modified Diebold-Mariano test.

I find that there exists an improvement in the forecast accuracy of machine learning models over a linear regression. Out of all tested models, the Light Gradient Boosted Machine (LGBM) model demonstrated superior performance most of the time for all targets. The dollar volume bars sampling showed to yield a more accurate and stable out-of-sample model performance than the time bars sampling. Overall, the market was found to be efficient both in normal times and in crisis times.

These results suggest that machine learning techniques can lead to improvements in our abilities to predict market microstructure variables. The ability of machine learning to process complex data using non-parametric algorithms makes it easier to extract patterns from the data, previously unrecognizable by the parametric models.

The remainder of the paper is organized as follows. In Section 3.2 I discuss the existing literature on the topics of Machine learning for Big Data and forecasting, and its applications to market microstructure. Section 3.3 outlines the methodology of the current paper, the machine learning methods I use, while section 3.4 describes the data and the microstructure measures. The empirical results are presented in section 3.5. Finally, section 3.6 concludes and discusses the possibilities for future research.

3.2 Literature Review

This paper builds on the previous work from several strains of the literature: machine learning and big data analysis for economic research, machine learning for market microstructure, and market stability and price prediction during periods of stress. I outline the main findings and methods within those topics in the following paragraphs.

The first strain of the literature, is the general application of machine learning and big data analysis for economic and financial research. Machine learning and big data are used more and more often for economic research due to the availability of the data and the promising

results from the other fields². Machine learning techniques can be successfully applied to analyze economic problems that embed big data. For example, Cavallo and Rigobon (2016) rely on social media to collect massive amounts of data on retail prices from more than 50 countries. They focus on measurement rather than prediction to study price stickiness and to check whether the “one price law” concept of the international economics holds. The most recent research that builds on the features of big data is summarized in Goldstein et al. (2021).

While machine learning techniques are relatively easy to use nowadays, it is worth noting that to apply them correctly requires finding relevant tasks. Varian (2014) argues that conventional econometric tools may not address the issues that are unique to big datasets, and proposes to use machine learning to model complex relationships. As pointed out by Mulinathan and Spiess (2017), machine learning is aimed mainly at the problem of prediction rather than parameter estimation, which is more common in economics. Bajari et al. (2015) apply LASSO regression, support vector machines, and other techniques to improve out-of-sample prediction accuracy in the problem of demand estimation. They show that predictions obtained via the machine learning approach are more accurate than some commonly used alternatives.

The recent abundance of big data in economics and finance rises new questions and provides the possibility of the future research in diverse directions like the adjustment of the corporations’ strategies by targeting the response from computers as decision-makers (P. Bond et al. (2012), Cao et al. (2020)), positive and negative impacts of big data on financial markets (Chawla et al. (2016), Shiller (2015)), possibility to measure previously unobserved economic activities by processing more complex data (Gentzkow et al. (2019), Li et al. (2021), Gerken and Painter (2019)), adjusting the financial market regulation due to the presence of computer algorithms (O’hara et al. (2014), Spatt (2020)), and others.

This paper also relates to the literature on the applications of machine learning in market microstructure. In the current age of fast markets, the structure of the markets’ functioning plays a crucial role in explaining market behavior. As was shown by Philip (2020), due to the complexity and non-linearity of microstructure of financial markets, commonly used techniques that do not address these issues may produce conflicting and incorrect inferences. That is why machine learning has gained recent popularity. For example, Rossi (2018) employs boosted

²For an overview of the ML techniques used in biology, physical sciences, biomedicine, and material sciences, see respectively Tarca et al. (2007), Carleo et al. (2019), Wainberg et al. (2018), Wei et al. (2019)

regression trees to predict monthly returns and volatility. This method delivers a greater out-of-sample predictive accuracy, compared to the standard models, and suggests a non-linear relation between the optimal portfolio allocation and the predictor variables. Krauss et al. (2017) already use daily returns to train deep neural networks, gradient-boosted-trees, and random forests in the context of statistical arbitrage on the S&P 500. The promising results obtained by the authors allow them to challenge the semi-strong hypothesis for market efficiency.

This trend in market microstructure research that applies machine learning, to shift towards higher frequency time frames is pursued further by Chinco et al. (2019), who use LASSO regression to generate one-minute-ahead return forecasts that provide better fit out-of-sample due to capturing unexpected, short-lived and sparse predictors. Bartlett et al. (2022) use proprietary order book data to study the predictive power of the odd lot quotes submissions on the future prices at one-minute intervals. They use XGBoost machine learning prediction algorithm to create a profitable trading strategy that uses odd lot data.

As computers took over the dominant trading role, the high-frequency trading (HFT) impact on financial markets was also studied theoretically. When developing a dynamic limit order market model with asymmetric information and HFT, Arifovic et al. (2021) follow Chiarella et al. (2015) and employ a classifier system based on a genetic algorithm (GA). They use GA as an adaptive learning mechanism to study how changes in information and speed affects HFT. The advantageous part of using machine learning for such a model is that the limit order book is extremely high-dimensional, and the GA allows traders to learn from this state space.

Finally, the paper builds upon the literature on market quality and stability. In this context, the microstructure of the market plays an important role due to the fact that high-frequency trading (HFT) and algorithmic trading are now common market practices. Previous works such as Hendershott et al. (2011), Easley, López de Prado, and O'Hara (2012), Brogaard et al. (2018), and Gonçalves et al. (2021) study the impact of HFT on markets stability, quality, and liquidity in the presence of HFT.

This paper will try to shed more light on whether microstructure information such as order flows in the limit order book, short-run price movements, and other information, that is usually used by HFT is available to predict future markets liquidity, volatility, and the probability for the markets to crash (experience an extreme price event). The predictive power of microstructure metrics will be assessed by comparing the out-of-sample forecast performance

of the standard linear model and highly non-linear models from the machine learning toolbox. Moreover, the paper will try to improve the work of Easley et al. (2021) and assess the performance of machine learning techniques in forecasting continuous variables.

3.3 Methodology

The main aim of the current paper is to assess predictability using diverse microstructure variables.

I use machine learning models, to study the ability to forecast some important liquidity, volatility, and market stability measures by using a collection of general market microstructure variables. Additionally, I compare the forecasting accuracy of the models during normal time versus crisis time by capturing the outbreak of the COVID-19 crisis in the data sample.

3.3.1 Linear model

The prediction accuracy of all models described below is compared to the most common benchmark, the simple linear predictive regression, estimated via ordinary least squares (OLS). Linear regression is not expected to perform better for the current high-dimension forecasting exercise, and is used mainly as a reference model to highlight the features of more complex and non-linear methods.

Linear regression model requires that a linear function of a vector of parameters and the set of predictor variables can approximate the set of conditional expectations $g(\cdot)$

$$g(X_{i,t}; \theta) = X'_{i,t} \theta. \quad (3.1)$$

Due to the quite simple specification, no nonlinearity is allowed in such a model. The interactions between the regressors are not taken into account either.

The objective function, for the estimation of this model,

$$\mathcal{L}(\theta) = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T (y_{i,t+1} - g(X_{i,t}; \theta))^2, \quad (3.2)$$

gives the OLS estimator, when $\mathcal{L}(\theta)$ is minimized. The advantage of this l_2 objective function is

that it offers a simple analytical form for the estimates without complex optimization algorithms involved.

The objective function of a simple linear regression may be extended to improve the robustness of the model, or to account for the heavy tails in the distribution of the independent variable (see Gu et al. (2020)). In case when the number of regressors approaches the sample size, the linear model becomes at best inefficient. Various penalization methods, from LASSO regression to the Elastic Net, are used to tackle the problem of efficient and consistent model estimation when the signal-to-noise ratio is low. However, this is not the case in the current paper, so the objective function of the model was not altered.

It is worth noting that the linear regression in general case is not a “subset” of machine learning methods. For example, one can use a neural network with zero hidden layers and a linear activation function to obtain an analog of the linear regression in the machine learning context. However, any non-redundant architecture of a machine learning model would aim to identify non-linear relationships in the data. When the size of the dataset is not sufficient to train a complex model with a large number of parameters, the simple linear regression may yield both more stable and accurate inference or predictions. That is why the most appropriate model to use is often identified on a case by case basis and usually after an initial data analysis.

3.3.2 Light Gradient Boosted Machine (LGBM)

One alternative way to account for some complex interactions between predictors is to employ regression trees. This is a popular approach that is fully nonparametric and is being used more and more often in economic research. The trees form clusters of observations that are similar to each other. The so-called “branches” of the tree are growing at each step of the algorithm, that splits the predictors into groups by minimizing an error function built on the features and targets in the training data set. In order to grow the tree, the algorithm has to find partitions that best differentiate between the potential outcomes.

Previous research like Breiman et al. (2017) has developed several optimization techniques that make the algorithm converge on an optimal tree. Nevertheless, regression trees can overfit the data quite often. This drawback is usually solved by regularization. One can consider an ensemble of trees and combine the forecasts from many trees into one.

For this paper, I use the method of “boosting” to combine many trees into a single one. The

method is based on the idea that many poor performing individual trees, when combined, can yield a more accurate and a more stable model than a single but complex tree. A simplified algorithm of gradient boosting can be described as follows. In the first step, one simple tree with a small depth (L) is created. This tree does not provide any good predicting accuracy, and is biased. In the second step, another shallow tree is created, that is designed to fit the forecasting residuals from the first tree. The predictions of both trees are combined into one by shrinking the forecast from the second tree by a predefined factor $\nu \in (0, 1)$. At every next step, the new shallow tree is designed to fit the residuals from the combined model at the previous step, and its forecast is shrunk to prevent overfitting and added to the ensemble. The parameters L, ν, N of the final model of N shallow trees are chosen based on a cross-validation procedure.

XGBoost which stands for an extreme gradient boosting algorithm is a highly scalable and flexible tool that is able to improve over other boosting techniques due to a faster and more robust regularization method. A light gradient boosting machine (LGBM) is an implementation of the XGBoost by Microsoft team with reduced convergence time. Also, LGBM is usually more accurate than XGBoost, and is better suited for handling big data. The difference comes mainly from the way the trees are grown in the model. In LGBM the trees are growing by increasing the size of the current leaf first, instead of checking all of the previous leaves for each new leaf.

There are many hyperparameters in the LGBM one can tune to get a better forecasting performance. I will focus on just one important parameter which is the number of decision trees used in the ensemble. Decision trees are usually added to the model sequentially in order to correct and improve the forecasts produced by previous trees. The simple intuition then suggests, that the more trees the model has the better the forecast will be. I will fit LGBM models with 50, 100, and 200 trees in the ensemble³. Therefore, I will study the performance of the LGBM(50) through LGBM(200) models in forecasting microstructure variables during normal and crisis times and compare the accuracy of the predictions coming from LGBM with Neural Network models, as well as with the benchmark linear regression.

³LightGBM package in Python sets 100 trees as the default number of trees for LGBM models.

3.3.3 Neural Networks

Another non-linear method I will use to analyze the forecasting performance of the microstructure measures is the artificial neural network. This method is currently the top choice for such machine learning problems as natural language processing, computer vision, and others. Neural Networks are extremely flexible due to employing a set of layers of nonlinear interactions between regressors. This set of entwined layers can vary from relatively small (shallow networks) to quite large (deep networks). Due to potentially high complexity of neural networks, this method is considered the least transparent, highly parameterized, and almost uninterpretable among other machine learning tools.

Shallow Networks

For the shallow network, when the depth of the book is small, the analysis focuses on the so-called “feed-forward” networks. As figure 3.1 illustrates, these networks consist usually of an input layer, hidden layers, and an output layer.

Figure 3.1 around here

The input layer represents raw predictors that, together with some bias, are passed to the first hidden layer. Hidden layer(s) transform the signal from the predictors in a non-linear way, and the resulting outcome is aggregated at the output layer. The dimension of the input layer equals the dimension of the predictors. A signal from every node (or neuron) is scaled according to a parameter vector θ , that usually includes weight parameters, and propagated linearly to the next layer. Then, every neuron applies a nonlinear function, which is also called “activation function”, to the aggregated inputs. For example, the first neuron in the first hidden layer transforms inputs from the predictors as $h_1^{(1)} = f\left(x_0 + \sum_{i=1}^3 x_i \theta_{1,i}^{(0)}\right)$, where x_0 is a bias term. Finally, the outputs from the last hidden layer are aggregated linearly into an ultimate output:

$$\hat{y}_1 = h_0^{(2)} + \sum_{i=1}^3 h_i^{(2)} \theta_i^{(2)}. \quad (3.3)$$

There are numerous options for choosing the nonlinear activation function for the neurons. There exist sigmoid, softmax, hyperbolic, and many other functions. However, for this paper, for the activation of hidden layers, I use another popular function called rectified linear unit,

or ReLU for short. The ReLU function is defined as follows

$$\text{ReLU}(x) = \begin{cases} 0 & \text{if } x < 0, \\ x & \text{otherwise} \end{cases}, \quad (3.4)$$

and a simple linear function for the output layer.

The neural network weight parameters are estimated by minimizing the objective function of prediction errors. The gradients of the cost function are calculated by propagating the error in the output back through the network. With the error at each node calculated, the amount that each node contributes to the overall cost can be estimated, leading to the gradient.

Selecting a network architecture is a complicated task. There is an infinite number of architectures to choose from, but only a finite amount of time to find the optimal one. Instead of checking all of them, I consider architectures up to four hidden layers for forecast generation and compare their performance. The shallowest network (NN1) will consist of just one hidden layer of 16 neurons, and the network with four hidden layers (NN4) has the following configuration of neurons in it: 16, 8, 4, and 2 neurons at levels one to four, respectively.

Deep & Recurrent networks

Eldan and Shamir (2016) show that an increase in the depth of the neural network can be substantially more beneficial than increasing the number of neurons in each layer. Around these times, deep and recurrent⁴ networks gained more popularity within the machine learning community. It is difficult, however, to train recurrent or very deep networks due to the fact that they are prone to the vanishing gradient problem. Long-short term memory (LSTM) was introduced by Hochreiter and Schmidhuber (1997) as a powerful recurrent neural network that was designed to be an efficient tool to overcome exploding or vanishing gradient issues.

The LSTM block architecture is shown in Figure 3.2. This figure shows the main components of an LSTM, among which are a cell, an input gate (σ_1), a forget gate (σ_2), and an output gate (σ_3). The architecture also contains the input signal $x^{(t)}$, the output $y^{(t)}$, and the activation functions g_1 and g_2 that are usually realized in the form of a hyperbolic tangent. The output of the block is then connected to the next block's input.

⁴In the recurrent network the outputs from some or all hidden layers become inputs for previous hidden layers, thus creating the recurrent structure.

Figure 3.2 around here

The cell remembers values over some time interval, and the gates regulate the flow of information linked to the cell. The gates are represented by sigmoid activation functions. The input gate combines the current input and the output of the LSTM unit in the last iteration. The forget gate determines which information should be removed from the previous cell states of the LSTM model. The output gate combines the current input and the output of the LSTM unit in order to calculate the current output.

Even though there is research that tries to improve the performance of the model (see, for example, Bayer et al. (2009), Bellec et al. (2018), and Su and Kuo (2019)), the vanilla LSTM performs very well. Fischer and Krauss (2018) apply LSTM model to financial markets prediction and show that LSTM outperforms logistic regression, standard deep neural network, and the random forest approaches. This suggests that LSTM can perform well when the data is non-linear, non-stationary, and sequence-correlated.

Therefore, I will study the performance of the LSTM in forecasting microstructure variables during normal and crisis times and compare the accuracy of the predictions coming from LSTM with other models. For comparison, I will estimate the parameters of the network where the hidden layer has 2, 4, 8, and 16 LSTM blocks or neurons with the default sigmoid activation function.

3.3.4 Accuracy metrics for performance comparison

Although the models are trained using the mean squared error loss function, to evaluate their out-of-sample performance, I calculate the out-of-sample R^2 as follows

$$R_{\text{oos}}^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum_{t \in \mathcal{T}^{\text{test}}} (y_t - \hat{y}_t)^2}{\sum_{t \in \mathcal{T}^{\text{test}}} (y_t - \bar{y})^2}, \quad (3.5)$$

where RSS and TSS stand for, respectively, the residual and the total sum of squares out-of-sample, $\mathcal{T}^{\text{test}}$ is a set of out-of-sample observations, that were not exposed to the model during the training procedure, y_t is a target variable the model is aiming to predict, \bar{y} is its mean, and \hat{y}_t is the model output at time t .

The out-of-sample R^2 is also used for identifying the relative importance of individual features for every model. The feature importance can be estimated by calculating the reduction

in R_{oos}^2 from setting all values of the given feature to a constant within the training sample. Obtained feature importance is then normalized within the model, to sum up to one, that allows for interpreting it as relative importance for the model.

When comparing the performance of several models out-of-sample, selecting the model with the highest R_{oos}^2 is not enough. One has to determine, whether the difference in their performances is statistically significant or comes from the specific properties of the selected sample. To analyze and compare the models, I use Diebold and Mariano (1995) test with modifications suggested by Harvey et al. (1997) for differences in out-of-sample predictive accuracy.

The basic concept of the modified test can be summarized as follows. Suppose, that the difference between the first set of predictions of size T and the true values is e_1 , and between the second set of predictions of the same size T and the true values is e_2 . Then, the loss-differential is given by

$$d = e_1^2 - e_2^2. \quad (3.6)$$

The null hypothesis is $E(d) = 0$, and the test statistics follow the student-t distribution with $T - 1$ degrees of freedom⁵. The obtained p -values from this statistic will be presented during the model comparison.

3.4 Data

In this section, I describe data sources, construction of time and volume bars, and construction of features that will serve as inputs to machine learning algorithms described in the previous section.

The data for this study come from Nasdaq's historical ITCH in the form of the order book messages. Traditionally, market microstructure analysis employs the Trade and Quote (TAQ) data. For the purpose of the study, the order book message data has a considerable advantage relative to the TAQ data but also one limitation. The advantage is that the order book message data are time-stamped in nanoseconds order submissions, executions, cancellations, and modifications on the Nasdaq equity market. This type of data allows obtaining complete order book information, whereas from TAQ data one can get only top-of-book quotes. Order book message data, therefore, allow for a precise realized volatility computation as well as for

⁵For more information on how the statistic is derived and how it behaves, please refer to Harvey et al. (1997).

direct observation of liquidity provision on each depth level of the limit order book at any time.

The limitation of the data, used for this research, is that it provides limit order book information from Nasdaq exchange only. The order book activity on other venues may be different. Trades on Nasdaq account for, on average, 33% of trading activity for Nasdaq listed stocks, about 12.5% for NYSE stocks, and 16% for ARCA stocks. Despite the high fragmentation nature of financial markets, we share the reasoning of Brogaard et al. (2018) that liquidity transfers to other venues are unlikely due to the short period of interest and overall similar liquidity provision rules among exchanges. Thus, we argue that although the results obtained could not readily expand to other exchanges, still they should be taken into account.

The sample period encompasses two years of trading from January 2, 2019, to December 31, 2020. This period covers trading in the limit order markets as usual and also trading during a crisis period⁶. I consider only large-cap constituents of the set of firms traded on Nasdaq, where the trading activity is sufficiently high⁷. In total, the sample contains data on 524 liquid stocks that are traded on Nasdaq during market hours.

As was mentioned previously, the data comes in a series of timestamped messages. Therefore, before it is ready to be sampled into time and volume bars, it has to be processed. In the first step, it is necessary to connect any initial order to the modified versions of this order, as they have different order reference number. Then, the time series of prices and limit order book properties such as spread, depth, and others can be produced at the required frequency.

3.4.1 Time and dollar-volume bars

For the current study, the nanosecond frequency data is aggregated into bars based on clock time and dollar volume traded. Academic research employs extensively data aggregation based on either clock time, trade time, or volume. Easley, M. M. L. De Prado, et al. (2012) argue that low frequency traders are not completely defenseless against HFTs if they adopt volume-clock paradigm, Easley et al. (2016) also find bulk volume classification to be a useful tool to discern trading intentions from market data. Barardehi et al. (2019) propose to use trade time liquidity measures for microstructure research in order to obtain more accurate proxies for trading costs and to better explain the cross-section of returns. According to M. L. De Prado (2018), using

⁶The outbreak of COVID-19 crisis began in the end of February, 2020.

⁷A stock is considered to be sufficiently liquid if there are at least 3,000 trade messages for this stock during the trading day.

dollar-volume bars can achieve normalization of the dollar value transacted across time periods, thus, providing more stability to the sampling frequency.

I follow Easley et al. (2021) and use both dollar-volume bars and clock time bars sampling to compare the results of the machine learning exercise. The bars are formed as follows: the n -th bar is filled at the earliest time t_n that satisfies the condition

$$\sum_{t=t_{n-1}}^{t_n} p_t \cdot V_t \geq L, \quad (3.7)$$

where p_t is the price of the trade at time t , V_t is the number of shares traded at time t , t_{n-1} is the time when the previous bar was filled, L is the dollar amount of the volume bar. This amount is chosen such that during the trading day one has around 195 dollar-volume bars⁸. It is different, for each stock, since the dollar trading volume is different across stocks in the sample. However, the number of bars during the day is roughly the same for all stocks. If the trading activity in some stock is highly above (below) average, the dollar-volume bars will fill faster (slower) and there might be more (less) than 195 bars a day.

For the time bars, for each stock, I split the trading day into 2-minute intervals and then compute the microstructure features to serve as inputs and output targets for the machine learning models. The continuous trading period on Nasdaq lasts for 6.5 hours, from 9:30 a.m. to 4 p.m., which results in 195 2-minute intervals. This number, for the ease of the results comparison, is consistent with the average number of dollar-volume bars of the alternative sampling.

3.4.2 Features and targets construction

To assess the predictability of financial markets through machine learning, I employ a wide range of microstructure measures. These measures may serve as features (inputs) and also as targets (outputs) for models interchangeably. Below, I discuss the measures in order of increasing construction complexity.

One of the most intuitive measures is the volume-weighted average price during either the 2-minute interval or the dollar-volume bar. It is accompanied by the short and long horizon returns. For the short horizon return I use a look-back window of 10 periods, and for the long

⁸The value of L , for each year, is chosen based on the average daily trading volume in the previous year.

horizon a window of 50 periods. An additional feature, that is easily observed, is the total depth of the market 50 basis points around the midquote at the end of each period or volume bar. It is useful, to record the absolute values of the dollar amounts that are required to immediately move the price 50 basis points up or down. This measure will combine both, the depth and the width of the market around the midquote and can represent Up/Down market resistance. I also employ, some naturally coming to mind microstructure features, like realized volatility during the interval, quoted spread in the limit order book, and realized spread, for which the look-back window of 10 periods is used.

The order book message data contains various messages that come from different market participants. A message may instruct to add a new order or to delete, modify, execute, or partially execute the existing one⁹. Easley, López de Prado, and O’Hara (2012) and Chiarella et al. (2015) show that the order flow may be an important predictor in the current high-frequency markets. That is why it is natural to measure the intensity flow of different orders within each period/bar.

Also, some market participants may share their ID with the exchange when submitting an order. There are several types of Nasdaq participants, among which are Market Maker (M), Exchange (E), Order Entry Firm (O), Electronic Communication Network (E), and others. The share of new order submissions, with market participant ID provided, is around 10% of all orders. I, therefore, do not distinguish between different market participants, but simply record (i) the fraction of orders within the interval that had a market participant ID provided, (ii) the fraction of shares added to the limit order book by identified market participants, (iii) the fraction of dollar volume, and (iv) the weighted (by distance from the midquote) presence in the limit order book of the identified market participants

$$\text{Weighted Presence}_{\text{MPID}, t_n} = \sum_{k=1}^K \frac{V_{k, t_n}^{\text{MPID}}}{\left(k - \frac{1}{2}\right) \tau} \bigg/ \sum_{k=1}^K \frac{V_{k, t_n}}{\left(k - \frac{1}{2}\right) \tau}, \quad (3.8)$$

where V_{k, t_n} is available shares volume, located $\left(k - \frac{1}{2}\right)$ price ticks τ away from the midquote, at time t_n , and K is some integer, big enough to cover all levels on the price grid with shares available. For more details on the above measure, refer to the Appendix 3.A.2.

A more sophisticated measure, that is based on the limit order book shape, is the depth

⁹For a full list of message types, refer to the Appendix 3.A.1.

imbalance. I distinguish two types of depth imbalances (i) unweighted imbalance, that is simply the difference between available volume to buy and volume to sell within some distance around the midquote ¹⁰, and (ii) weighted depth imbalance where the available volume, at every price level, is scaled based on the distance from that price level to the midquote. I believe, it is important to complement the unweighted depth imbalance with the weighted one, because the shape of the limit order book is an important piece of information for generating profitable trading strategies by HFTs.

As HFTs play an important role in the processes of price formation in the current high-frequency markets, it is reasonable to measure the HFT activity during the day. The proxy for the HFT activity is constructed similarly to a measure proposed by Hasbrouck and Saar (2013) and later applied by Gonçalves et al. (2021). This measure calculates the intensity of “strategic runs”, which are series of linked messages. The linking results from HFT dynamically submitting and canceling orders to incorporate the latest information into prices.

Following their methodology, we connect a newly submitted limit order to a previously deleted order if the time between the two events does not exceed 100 milliseconds. The newly submitted order should have the same direction and size in shares as the previously deleted one. Only sufficiently long runs of 10 and more linked orders are kept. We scale the sum of durations of all runs, which are allowed to overlap, by the duration of the time or volume bar. The proxy for the HFT presence for bar t_n is defined as follows:

$$HFT_{t_n} = \frac{1}{t_n - t_{n-1}} \sum_{j=1}^N T_j, \quad (3.9)$$

where N is the number of strategic runs during the bar t_n , and T_j is the duration in seconds of run j .

Additionally, I identify whether a time (or dollar-volume) bar contains a mini-flash crash. Following Gonçalves et al. (2021), I compute extreme returns for each bar and identify the intervals containing mini-flash crashes as those where the Z -score for the midquote extreme returns exceeds the critical value. The procedure gets a total of 32,681 mini-flash crashes for sampling based on time bars, and 54,408 mini-flash crashes for the sampling based on dollar-volume bars. This difference between the number of crashes identified depending on the

¹⁰The distance is chosen to be 50 basis points around the midquote.

sampling type, may support the arguments for using dollar-volume bars sampling as it captures more microstructure features of the data.

Lastly, I employ some well-known microstructure measures like Roll measure, Amihud illiquidity measure, and volume synchronized probability of informed trading (VPIN) and the Variance ratio (VR). However, according to Easley et al. (2021), these measures were initially defined to be computed based on the low frequency data (daily frequency at the minimum). In order to adapt those measures to the current high-frequency data, I apply the concept of time periods and dollar-volume bars to the original definition.

In particular, Roll measure is defined as

$$R_{t_n} = 2\sqrt{|\text{Cov}(\Delta P_{t_n}, \Delta P_{t_{n-1}})|}, \quad (3.10)$$

where ΔP_{t_n} is a vector of differences in the close price between bars t_{n-1} and t_n .

Amihud measure is defined as

$$\lambda_{t_n} = \frac{1}{W} \sum_{i=t_n-W+1}^{t_n} \frac{|r_i|}{p_i V_i}, \quad (3.11)$$

where V_i , p_i , r_i are respectively volume, price, and return for the bar i , and W is the look-back window.

Variance ratio, according to Lo and MacKinlay (1989), is estimated as

$$VR(5) = \frac{\text{Var}(r_{5 \cdot W})}{5 \cdot \text{Var}(r_W)}, \quad (3.12)$$

where r is the return, and W is the look-back window.

Finally, the volume-synchronized probability of the informed trading is computed as follows

$$\text{VPIN}_{t_n} = \frac{1}{W} \sum_{i=t_n-W+1}^{t_n} \frac{|V_i^B - V_i^S|}{V_i}, \quad (3.13)$$

where V^S is the volume of the trade that was initiated by a seller, and V^B was initiated by a buyer, and W is the look-back window.

The final list with more description of features and targets can be found in the Appendix 3.A.2. Tables 3.1 – 3.4 present some summary statistics of the distribution of different measures

throughout the sample. The results are presented per year, per sampling method.

Tables 3.1 – 3.4 around here

When examining the summary statistics for both years, one can note that the crisis period is characterized by higher quoted spreads but lower realized spreads, higher HFT participation, lower participation from the identified market agents. When the markets become more volatile, e.g. during periods of crisis, HFTs' advantage in speed may make their trading strategies more profitable, while the low-frequency traders are worse off and reduce their participation in the market. The probability of a mini-flash crash stayed roughly the same. This fact suggests that mini-flash crash occurrence is not directly linked to the economic situation in general.

The depth imbalance measure became much more volatile in 2020 with time bar approach but the effect the opposite for the dollar-volume bars based approach. Also, sampling using dollar-volume bars identified more flash crashes, higher intensity of the order flow, and higher participation of the HFTs. In general, dollar-volume bars provide us with more stable microstructure features.

3.5 Empirical Results

In this section, I present the results of the model fitting and forecasting exercise carried out for the set of different microstructure variables (or targets) such as (i) returns, (ii) mini-flash crash occurrence, (iii) quoted spread in the limit order book, and (iv) realized volatility. The analysis uses a standard open source software on machine learning, Scikit-learn (Pedregosa et al. (2011)), Keras (Chollet et al. (2015)), and LightGBM packages in Python. The targets are chosen in such a way, in order to cover the most intriguing aspects of the market microstructure.

Returns is the most common metric that is widely used in other areas of economic research but also being the most desirable to forecast. The flash crashes occurrence is important as it indicates the general market stability and the presence of inefficiencies in the market. The quoted spread is the main liquidity measure that many trading strategies are built upon. It is also important in the context of informed and insider trading as the spread and the presence of informed traders tend to correlate. Finally, the realized volatility measure, that was shown to best estimate market volatility in the high frequency environment, is used as a target for

the forecast. In the following subsections, I discuss models' performance for each target and compare differences in forecast accuracies between (i) sampling methods and (ii) calm and crisis times.

3.5.1 Returns

Returns prediction in a short term is generally considered a difficult task. The strong form of the efficient markets hypothesis states that all information about the future direction of the price is already incorporated into the current price, therefore nothing should give a positive edge in forecasting the short-term returns. Clearly, the reality does not reflect the theory perfectly, nevertheless, the results of the forecasting exercise show that financial markets, in aggregate, were quite efficient in 2019 and 2020.

The first column in Table 3.5 presents the 2019 and 2020 out-of-sample R^2 for the forecasts of one-period returns for twelve models with time bars sampling of the data. With the length of each time bar of two minutes, the data contain more than 15 (13) million observations in the training set and more than 1.6 (1.5) million observations in the testing set in the year 2019 (2020). The results of the model fitting are based on the pooled data. In this setting, the reference number of the stock is also employed by the models as a factor variable. The results from the individual regressions for returns and all the other targets coincide with the pooled one, but are, in general, more dispersed.

The out-of-sample R^2 when predicting returns is remarkably low for all models. At the best, machine learning algorithms perform as good as naive mean prediction where the mean of returns fluctuates around zero. The situation is similar when I use the dollar volume bars to sample the data. Table 3.6 in the first column shows the R_{oos}^2 for return forecast in the next bar. Although LGBM models show higher R^2 values, the economical effect of this improvement is too low to be noticed.

To identify, if there are benefits of employing one model over another one in forecasting returns, I use the modified Diebold-Mariano test. Tables 3.7 and 3.11 show the statistical significance of differences among models at the time bars and dollar volume bars sampling. These tables report the test statistics for the pairwise comparison of a column model versus a row model. Modified Diebold-Mariano test statistics under the null hypothesis of no difference between the models are distributed according to a student- t distribution with $T - 1$ degrees of

freedom, where T is the length of the forecast series. The sign convention suggests that the positive value of the statistics indicates that the column model outperforms the row model. Bold numbers denote significance at a 5% level for each individual test.

Table 3.7 around here

The results of models comparison for the time bar sampling suggest that the simple linear regression in 2019 is not outperformed by any model except the LGBM model with 50 decision trees. The same statement can also be done about the set of LGBM and Neural network models. Contrary to those, LSTM models show worse performance compared to almost every model.

In 2020, however, the difficult choice of the best model resolves with LGBM model with 200 decision trees, LGBM(200) statistically outperforming every other model. At the same time, the linear model is outperforming both LSTM and Neural Network models. LSTM models in 2020 are statistically the worst in forecasting returns as they lose to almost all other models. Perhaps, a deeper LSTM would perform better, as it is possible to see from the comparison between LSTM(1) through LSTM(4).

When we look at the similar comparison in Table 3.11 that is based on the dollar volume bars sampling, we can see similar results. In 2019 it is uncertain, which models perform the best as there is no significant difference between the linear model and the LGBM, and as for the time bars sampling, the LSTM model is performing poorly. In 2020 we again observe a clear dominance of LGBM models over the others. Notably, the NN(4) is statistically worse than any other model, which can be a result of overfitting the training data.

Table 3.11 around here

3.5.2 Mini-flash crashes

As Gonçalves et al. (2021) suggest, the average duration of the mini-flash crashes identified from January 2018 to October 2020 is 47.9 seconds. If one confronts this with the average volume traded during the crash, it will be clear that time bars sampling may not be the best choice of data treatment when the aim is to identify and forecast such rapid but detrimental to the market events. The second column in Table 3.5 shows the out-of-sample R^2 of the forecasts of the flash crash events.

As the target variable is binary, the forecasted value can be treated as the probability of the mini-flash crash occurring in the next bar. In this case, linear probability regression is not the most natural choice (since the probability can be potentially negative or greater than one), but a very simple in implementation method. Naturally, we observe a much higher R_{os}^2 for machine learning methods compared to the linear regression who provides about 6%. The highest R^2 of 49.6% deliver the LGBM models in 2019, Neural network models provide around 26%, and LSTM models 13%. In 2020, the out-of-sample forecasting performance of all models stayed relatively stable.

Table 3.6, column 2, confirms that dollar volume bars provide a better specification for the data to predict, and possibly explain, mini-flash crashes. Even linear regression provides from 46% to 51% of the variability in the future values captured in the model output. Machine learning methods provide a significant increase in the prediction accuracy with R_{os}^2 improving up to 84.7% for LGBM(50) in 2020. Other models deliver on average around 60% in R^2 metric.

Tables 3.8 and 3.12 show the statistical significance of differences among models at the time bars and dollar volume bars sampling.

Table 3.8 around here

The results of models comparison for the time bar sampling suggest that the simple linear regression is statistically outperformed by every machine learning method both in 2019 and 2020. Unsurprisingly, LGBM models outperform other competitors by a lot. Also, from Table 3.8, we can observe a relatively poor performance of LSTM models. Most of the Neural network architectures were able to outperform the LSTM, except for the Neural network with one hidden layer. For the case of mini-flash crashes, we have stable results in the model comparison between 2019 and 2020, probably due to the similar amount of crashes detected within time bars in these years.

When we look at the similar comparison in Table 3.12 that is based on the dollar volume bars sampling, we can see similar results. LGBM models show the best performance compared to all other models, LSTM models are getting better with the increased number of modules, but still perform worse than all the other machine learning models, and Neural networks also require a deeper structure to be competitive with the LGBM.

Table 3.12 around here

3.5.3 Quoted spread

Quoted spread is defined as the difference between the best ask and the best bid prices in the limit order book, measured in basis points. Prediction of the spread is valuable for the design of trading and execution strategies, risk management, and for the general overview of market liquidity.

Column 3 of Table 3.5 shows that the forecast accuracy of all models when predicting quoted spread is considerably high. The average out-of-sample R^2 in 2019 exceeds 60% while the maximum value of 67% is delivered by the output of the LGBM(200) model. Interestingly, the predictive power of all models reduces dramatically in 2020. During the period of crisis, all models have difficulties providing a reliable forecast for the quoted spread since this measure is highly volatile during market downturns. The average R^2 delivered by all models drops to around 13% in 2020. Still, the LGBM model provides the best forecast according to the R^2 metric of 17.7%.

In contrast, the dollar volume bars provide a stable forecasting performance across the two years of estimation. Table 3.6 shows that the R^2 measure did not change drastically in 2020 compared to 2019. The overall performance of the models is slightly worse, however. The average R^2 is around 51% in 2019, and 45% in 2020. The best performing model based on the R^2 is LGBM(200).

When analyzing Diebold-Mariano test statistics for the pairwise comparisons of the models in the time bar sampling, Table 3.9 also suggests that LGBM(100) and LGBM(200) are the best performing models. In general, all LGBM models are superior, while the feed-forward Neural network models do not lose only to the linear regression. Among themselves, NN models with less number of hidden layers perform better. LSTM models strongly outperform the feed-forward networks, and in contrast, demonstrate better performance with the increase of the hidden LSTM modules.

Table 3.9 around here

In 2020, when the predictive accuracy of all models based on time bars falls, the linear regression happens to outperform Neural networks and LSTM models according to Diebold-Mariano test statistics. The LGBM models remain the best performing ones still. The other relations did not change much, and LSTM still provided a better forecast compared to the feed-forward

networks.

Table 3.13 provides pairwise comparison results of the models for the data sampled into dollar volume bars. For both 2019 and 2020, we can see similar results that can be summarized as follows: the linear regression shows poor performance compared to other models, LGBM models being top-performing models, and LSTM with many hidden layers outperforming the feed-forward neural networks.

Table 3.13 around here

3.5.4 Realized volatility

The realized volatility happened to be the easiest to predict microstructure measure in this exercise. Columns 4 in tables 3.5 and 3.6 show the out-of-sample R^2 measure for the forecast performance of the twelve models that have been estimated. The average R^2_{os} is about 95% for all modes, samplings, and years. There is however a couple of cases of quite low R^2 value in the Neural network models, but this is possibly due to the overfitting of the training data.

Table 3.10 suggests that LSTM models are the best-performing ones for the realized volatility when the data is sampled into time bars. LGBM models provide better forecasts in 2019 rather than in 2020 compared to the feed-forward neural networks, and the linear model sometimes can be as good as or even better choice than the sophisticated machine learning algorithms.

Table 3.10 around here

Table 3.14 provides pairwise comparison results of the models for the data sampled into dollar volume bars. In 2019, the difference between LGBM and LSTM models becomes most of the time insignificant or mixed. That year, LGBM models performance was also comparable to that one of the neural networks. LSTM model with three hidden modules is identified by the Diebold-Mariano statistics as the one that is either better than the other models in terms of the forecast accuracy or comparable to them. In 2020, LGBM models show again a very sound performance, however, the LSTM(4) is proven to give the most accurate forecast of the realized volatility in the next dollar volume bar.

Table 3.14 around here

3.5.5 The most important features

Further, I show the results of determining the relative importance of individual features for the forecasting performance for each target variable. This is done by calculating the reduction in the R_{oos}^2 from removing a certain feature from the training sample. For each target, the best-performing model was chosen to study the feature importance. Features importance within one model is normalized to sum to one, allowing for relative feature importance interpretation.

Returns

Figure 3.3 shows the microstructure features that are among the top 10 features in terms of relative importance for the returns forecasting. The four panels present feature importance analysis for dollar volume versus time bars and year 2019 versus year 2020.

Figure 3.3 around here

By all means, we observe the difference in the allocation and the ranking from year to year, but in general, the most important features for the dollar volume bars are the timestamp that is just a time of the day, current one-period return, realized volatility, weighted depth imbalance, roll measure, and the probability of the informed trading. These measures change rankings sporadically as an indication that returns forecasting in a short period is a challenging task. It seems reasonable, however, that liquidity and volatility features play a role in the future return realization.

For the time bars sampling, we can see similar features being relevant, but also one can see the influence of the order flow on the returns forecast. In particular, messages with an F flag are quite important for the time bars sampling. Those messages to the exchange indicate a new order submission by an identified market participant. The fact that this feature accounts for about 7.5% predictability in the returns suggests that there might be a way in associating the order submission intensity of the identified participants and future directions of the price.

Mini-flash crashes

Figure 3.4 shows the relative importance of the top 10 features for the prediction of the occurrence of the mini-flash crash in the next bar. As previously, the four panels present feature importance analysis for dollar volume versus time bars and year 2019 versus year 2020.

Figure 3.4 around here

For the dollar volume bars, the most important features remained at the top of the list throughout the whole period. The most significant one is the variance ratio that is usually referred to as a proxy for market efficiency. There might be the case that flash crashes occur when market inefficiency needs to be resolved, and therefore this market efficiency proxy contributes around 9% to the forecasting performance. Secondly, short- and long-term returns are relevant features, which possibly indicates that returns momentum has something to do with flash crash prediction. The realized volatility and the flash crash occurrence are meaningful because of the momentum effects as well. Finally, the locate code of the stock plays a role in the forecasting accuracy which might suggest that some stocks are more prone to flash crashes than others.

When examining feature importance for the time bars sampling, one can note the most relevant (10%) time feature. Indeed, previous research on flash crashes finds that such events tend to cluster in time. Market efficiency proxy is also an important feature. After the intuitive features like realized volatility and long-term returns, one can notice that the order flow and the number of executions and order submissions by the identified market participants also help in predicting a crash. Interestingly, neither for time bars nor for the volume bars sampling the HFT presence does not seem to help generate better forecasts for the flash crash occurrence. This fact may further support the claim that HFTs do not initiate the crashes themselves but mainly participate during the active phase of the crash by supplying liquidity.

Quoted spread

Feature importance for the quoted spread predictions is presented in Figure 3.5. One can note, that this target is more autocorrelated than the previous ones as the current value of the quoted spread helps predicting the future value. Price, however, is the most important predictor of the spread measure. This might be unintuitive for someone since the price is measured in dollars, but the spread is measured in basis points, but, at a second thought, it makes more sense higher price is usually correlated with a higher market capitalization of a company and higher capitalization brings more trading for the stock, and, therefore, more competition in the limit order book, that usually reduces the spread.

Figure 3.5 around here

Available depth in the limit order book, as well as the current realized volatility, naturally make good predictors for the spread liquidity measure. Also, in 2020, for both dollar volume and time bars, the order flow becomes an important indicator for the future spread, as at times of crisis, understanding the origination of the liquidity coming to the book is crucial. This is further strengthened by fact that the probability of informed trading becomes particularly important in 2020 for the volume bars. As informed trading is connected to the overall market liquidity, the fact that the level of informed trading helps predicting quoted spread in the next volume bar is not surprising.

Realized volatility

Finally, Figure 3.6 presents the most important features for the realized volatility predictions. It is evident that the realized volatility is highly persistent as the main predictive power of all models comes from including the previous period's realized volatility as a feature. On average, realized volatility accounts for more than 16% of the prediction accuracy.

Figure 3.6 around here

The timestamp is also an important feature as it is common knowledge that the volatility is usually higher during market opening and closing hours. For the dollar volume bars sampling, the illiquidity measures such as the Roll measure and the Amihud measure are relevant for the volatility prediction, while for the time bars sampling the intensity of order submission by the identified traders ends up being more important.

3.6 Conclusions and further directions

To conclude, I will summarize the results discussed above and outline the directions of further research.

Several machine learning approaches with different architectures were tested against each other and against the simple linear regression to find out which one provides more accurate forecasts of the market microstructure variables. Different targets for the forecasting were used that describe proxies for market volatility, stability, and liquidity. Out of all tested models, the Light Gradient Boosted Machine (LGBM) model demonstrated superior performance most

of the time for all targets. Long Short Term Memory (LSTM) neural network proved to be the best for the autocorrelated targets such as realized volatility, while a simple feed-forward neural network showed mixed results due to frequent overfitting training data.

The advantages of dollar volume bars sampling against time bars were also examined. The results show that dollar volume bars might work better for the prediction of variables for which the trading volume is an important component as the mini-flash crash example suggests. Moreover, dollar volume bars appear to be a more stable way of data sampling for the prediction of microstructure measures during times of market turbulence. The prediction accuracy of the quoted spread reduced dramatically in 2020 for the time bars but stayed stable for volume bars.

The question, of whether it is easier to predict the microstructure variables during crisis times is currently answered in the negative. The out-of-sample R^2 for the forecasts does not suggest worse performance in 2020 compared to 2019 for almost all models. For the quoted spread, the reduction in the prediction accuracy may be mitigated by using dollar volume bars sampling.

As this research is quite recent, there are still a lot of possibilities for improvement and directions to go further. One of the first ways to extend the current research, naturally coming to mind, is examining the cross-asset effects and examining the possible interconnections with the other markets like futures markets and options markets. The power of machine learning techniques is in the ease with which they handle complex nonlinearities between numerous elements in the global network. The financial market is an extremely complex system the network structure of which should be taken into account for forecasting exercises, but also for the purpose of inference.

Secondly, it would be beneficial to obtain a more granular view of market participants to possibly identify HFTs, market makers, informed, and other types of traders. This will potentially not only improve the forecasts but also will allow employing various agent-based models to further boost the prediction methodology that is supported by the theory.

Finally, one can go to even higher frequencies or study a longer period if the hardware allows.

References

- Arifovic, Jasmina, Xuezhong He, and Lijian Wei (2021). “Machine Learning and Speed in High-Frequency Trading”. *Available at SSRN 3866371*.
- Bajari, Patrick, Denis Nekipelov, Stephen P. Ryan, and Miaoyu Yang (May 2015). “Machine Learning Methods for Demand Estimation”. *American Economic Review* 105.5, pp. 481–85.
- Barardehi, Yashar H, Dan Bernhardt, and Ryan J Davies (2019). “Trade-time measures of liquidity”. *The Review of Financial Studies* 32.1, pp. 126–179.
- Bartlett, Robert P, Justin McCrary, and Maureen O’Hara (2022). “The Market Inside the Market: Odd-Lot Quotes”. *Available at SSRN 4027099*.
- Bayer, Justin, Daan Wierstra, Julian Togelius, and Jürgen Schmidhuber (2009). “Evolving memory cell structures for sequence learning”. *International conference on artificial neural networks*. Springer, pp. 755–764.
- Bellec, Guillaume, Darjan Salaj, Anand Subramoney, Robert Legenstein, and Wolfgang Maass (2018). “Long short-term memory and learning-to-learn in networks of spiking neurons”. *Advances in neural information processing systems* 31.
- Bond, Philip, Alex Edmans, and Itay Goldstein (2012). “The real effects of financial markets”. *Annu. Rev. Financ. Econ.* 4.1, pp. 339–360.
- Breiman, Leo, Jerome H Friedman, Richard A Olshen, and Charles J Stone (2017). *Classification and regression trees*. Routledge.
- Brogaard, Jonathan, Allen Carrion, Thibaut Moyaert, Ryan Riordan, Andriy Shkilko, and Konstantin Sokolov (2018). “High frequency trading and extreme price movements”. *Journal of Financial Economics* 128.2, pp. 253–265.
- Cao, Sean, Wei Jiang, Baozhong Yang, and Alan L Zhang (2020). “How to talk when a machine is listening: Corporate disclosure in the age of AI”. *NBER Working Paper*.
- Carleo, Giuseppe, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová (2019). “Machine learning and the physical sciences”. *Reviews of Modern Physics* 91.4, p. 045002.
- Cavallo, Alberto and Roberto Rigobon (May 2016). “The Billion Prices Project: Using Online Prices for Measurement and Research”. *Journal of Economic Perspectives* 30.2, pp. 151–78.

- Chawla, Nitesh, Zhi Da, Jian Xu, and Mao Ye (2016). “Information diffusion on social media: Does it affect trading, return, and liquidity?” *Return, and Liquidity*.
- Chiarella, Carl, Xue-Zhong He, and Lijian Wei (2015). “Learning, information processing and order submission in limit order markets”. *Journal of Economic Dynamics and Control* 61, pp. 245–268.
- Chinco, Alex, Adam D Clark-Joseph, and Mao Ye (2019). “Sparse signals in the cross-section of returns”. *The Journal of Finance* 74.1, pp. 449–492.
- Chollet, Francois et al. (2015). *Keras*. URL: <https://github.com/fchollet/keras>.
- De Prado, Marcos López (2018). *Advances in financial machine learning*. John Wiley & Sons.
- Diebold, Francis X and Robert S Mariano (1995). “Comparing predictive accuracy”. *Journal of Business & economic statistics* 20.1, pp. 134–144.
- Easley, David, Marcos M López De Prado, and Maureen O’Hara (2012). “The volume clock: Insights into the high-frequency paradigm”. *The Journal of Portfolio Management* 39.1, pp. 19–29.
- Easley, David, Marcos López de Prado, and Maureen O’Hara (2012). “Flow toxicity and liquidity in a high-frequency world”. *Review of Financial Studies* 25.5, pp. 1457–1493.
- Easley, David, Marcos López de Prado, Maureen O’Hara, and Zhibai Zhang (2021). “Microstructure in the machine age”. *The Review of Financial Studies* 34.7, pp. 3316–3363.
- Easley, David, Marcos López de Prado, and Maureen O’Hara (2016). “Discerning information from trade data”. *Journal of Financial Economics* 120.2, pp. 269–285.
- Eldan, Ronen and Ohad Shamir (2016). “The power of depth for feedforward neural networks”. *Conference on learning theory*. PMLR, pp. 907–940.
- Fischer, Thomas and Christopher Krauss (2018). “Deep learning with long short-term memory networks for financial market predictions”. *European Journal of Operational Research* 270.2, pp. 654–669.
- Gentzkow, Matthew, Bryan Kelly, and Matt Taddy (2019). “Text as data”. *Journal of Economic Literature* 57.3, pp. 535–74.
- Gerken, William Christopher and Marcus Painter (2019). “The Value of Differing Points of View: Evidence from Financial Analysts’ Geographic Diversity”. Available at SSRN 3479352.
- Goldstein, Itay, Chester S Spatt, and Mao Ye (2021). “Big data in finance”. *The Review of Financial Studies* 34.7, pp. 3213–3225.

- Gonçalves, Jorge, Roman Kräussl, and Vladimir Levin (2021). “Dark trading and financial markets stability”. *Working paper*.
- Gu, Shihao, Bryan Kelly, and Dacheng Xiu (2020). “Empirical asset pricing via machine learning”. *The Review of Financial Studies* 33.5, pp. 2223–2273.
- Harvey, David, Stephen Leybourne, and Paul Newbold (1997). “Testing the equality of prediction mean squared errors”. *International Journal of forecasting* 13.2, pp. 281–291.
- Hasbrouck, Joel and Gideon Saar (2013). “Low-latency trading”. *Journal of Financial Markets* 16.4, pp. 646–679.
- Hendershott, Terrence, Charles M Jones, and Albert J Menkveld (2011). “Does algorithmic trading improve liquidity?” *Journal of Finance* 66.1, pp. 1–33.
- Hochreiter, Sepp and Jürgen Schmidhuber (1997). “Long short-term memory”. *Neural computation* 9.8, pp. 1735–1780.
- Krauss, Christopher, Xuan Anh Do, and Nicolas Huck (2017). “Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500”. *European Journal of Operational Research* 259.2, pp. 689–702.
- Li, Kai, Feng Mai, Rui Shen, and Xinyan Yan (2021). “Measuring corporate culture using machine learning”. *The Review of Financial Studies* 34.7, pp. 3265–3315.
- Lo, Andrew W and A Craig MacKinlay (1989). “The size and power of the variance ratio test in finite samples: A Monte Carlo investigation”. *Journal of econometrics* 40.2, pp. 203–238.
- Mullainathan, Sendhil and Jann Spiess (2017). “Machine learning: an applied econometric approach”. *Journal of Economic Perspectives* 31.2, pp. 87–106.
- O’hara, Maureen, Chen Yao, and Mao Ye (2014). “What’s not there: Odd lots and market data”. *The Journal of Finance* 69.5, pp. 2199–2236.
- Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. (2011). “Scikit-learn: Machine learning in Python”. *Journal of machine learning research* 12.Oct, pp. 2825–2830.
- Philip, Richard (2020). “Estimating permanent price impact via machine learning”. *Journal of Econometrics* 215.2, pp. 414–449.
- Rossi, Alberto G (2018). “Predicting stock market returns with machine learning”. *Working paper*.

- Shiller, Robert J (2015). “Irrational exuberance”. *Irrational exuberance*. Princeton university press.
- Spatt, Chester S (2020). “Is equity market exchange structure anti-competitive”. *Working paper*.
- Su, Yuanhang and C-C Jay Kuo (2019). “On extended long short-term memory and dependent bidirectional recurrent neural network”. *Neurocomputing* 356, pp. 151–161.
- Tarca, Adi L, Vincent J Carey, Xue-wen Chen, Roberto Romero, and Sorin Drăghici (2007). “Machine learning and its applications to biology”. *PLoS computational biology* 3.6, e116.
- Varian, Hal R (2014). “Big data: New tricks for econometrics”. *Journal of Economic Perspectives* 28.2, pp. 3–28.
- Wainberg, Michael, Daniele Merico, Andrew Delong, and Brendan J Frey (2018). “Deep learning in biomedicine”. *Nature biotechnology* 36.9, pp. 829–838.
- Wei, Jing, Xuan Chu, Xiang-Yu Sun, Kun Xu, Hui-Xiong Deng, Jigen Chen, Zhongming Wei, and Ming Lei (2019). “Machine learning in materials science”. *InfoMat* 1.3, pp. 338–358.

Tables and Figures

Table 3.1. Descriptive statistics of the microstructure features in 2019, time bars sampling. The table presents mean, standard deviation, 25th, 50th and 75th percentile for the market microstructure measures calculated from time bars sampled data in 2019. The total number of observations is 17,176,544. The number of missing observations for each feature is also provided in the table. For the detailed description of the features, refer to the section 3.4.2.

Measure	Missing	Mean	StDev	25th	50th	75th
Price	0	112.212	211.837	38.950	65.165	126.130
Return (1)	86,011	$1.46 \cdot 10^{-6}$	$1.10 \cdot 10^{-3}$	$-3.42 \cdot 10^{-4}$	0.000	$3.48 \cdot 10^{-4}$
Return (5)	430,055	$5.09 \cdot 10^{-6}$	$2.40 \cdot 10^{-3}$	$-8.52 \cdot 10^{-4}$	0.000	$8.76 \cdot 10^{-4}$
Depth (\$M)	25,810	3.221	11.749	0.304	0.749	1.923
Up Resist. (\$M)	13,379	1.604	6.473	0.144	0.362	0.948
Down Resist. (\$M)	13,052	1.617	6.675	0.145	0.365	0.939
Realized Volatility	0	$2.43 \cdot 10^{-3}$	$2.31 \cdot 10^{-3}$	$1.24 \cdot 10^{-3}$	$1.87 \cdot 10^{-3}$	$2.89 \cdot 10^{-3}$
Quoted Spread	8,944	12.157	122.724	2.084	3.848	7.910
Realized Spread	544,652	-0.288	30.106	-12.898	-0.000	12.517
Intensity msg A	0	4.300	6.754	0.792	2.027	5.243
Intensity msg C	0	0.004	0.011	0.000	0.001	0.003
Intensity msg D	0	4.109	6.475	0.733	1.910	5.027
Intensity msg E	0	0.212	0.368	0.043	0.117	0.247
Intensity msg F	0	0.010	0.050	0.000	0.000	0.004
Intensity msg P	0	0.033	0.073	0.006	0.016	0.037
Intensity msg U	0	0.583	0.989	0.107	0.298	0.665
Intensity msg X	0	0.080	0.345	0.000	0.003	0.031
MPID orders	735	0.170	0.151	0.059	0.126	0.235
MPID shares	30,549	0.284	0.208	0.096	0.262	0.435
MPID \$ volume	30,549	0.715	0.279	0.526	0.833	0.946
MPID presence	30,930	0.055	0.086	0.011	0.027	0.062
Depth Imb.	184,196	-0.154	77.118	-2.725	-0.021	2.588
Depth Imb. (Wght.)	184,196	-0.002	0.276	-0.171	-0.001	0.167
HFT presence	0	0.024	0.058	0.003	0.008	0.021
Flash Crash	0	$9.61 \cdot 10^{-4}$	$3.10 \cdot 10^{-2}$	0.000	0.000	0.000
Roll	591,514	0.078	0.231	0.013	0.032	0.074
Amihud	211,886	$2.78 \cdot 10^{-8}$	$1.11 \cdot 10^{-6}$	$1.62 \cdot 10^{-9}$	$4.81 \cdot 10^{-9}$	$1.34 \cdot 10^{-8}$
Variance Ratio	806,902	0.816	143.480	0.328	0.549	0.903
VPIN	347,585	0.561	0.242	0.358	0.532	0.750

Table 3.2. Descriptive statistics of the microstructure features in 2020, time bars sampling. The table presents mean, standard deviation, 25th, 50th and 75th percentile for the market microstructure measures calculated from time bars sampled data in 2020. The total number of observations is 17,315,487. The number of missing observations for each feature is also provided in the table. For the detailed description of the features, refer to the section 3.4.2.

Measure	Missing	Mean	StDev	25th	50th	75th
Price	0	124.520	262.950	33.130	62.780	135.100
Return (1)	85,729	$-3.72 \cdot 10^{-6}$	$2.02 \cdot 10^{-3}$	$-5.38 \cdot 10^{-4}$	0.000	$5.34 \cdot 10^{-4}$
Return (5)	428,645	$-1.72 \cdot 10^{-5}$	$4.42 \cdot 10^{-3}$	$-1.34 \cdot 10^{-3}$	0.000	$1.33 \cdot 10^{-3}$
Depth	21,848	3.143	17.629	0.205	0.526	1.379
Up Resistance	11,621	1.544	9.618	0.097	0.255	0.674
Down Resistance	10,727	1.605	10.544	0.097	0.256	0.673
Realized Volatility	0	$4.18 \cdot 10^{-3}$	$4.67 \cdot 10^{-3}$	$1.82 \cdot 10^{-3}$	$2.97 \cdot 10^{-3}$	$5.02 \cdot 10^{-3}$
Quoted Spread	4,750	15.887	109.164	2.514	5.392	11.593
Realized Spread	376,663	-0.404	53.037	-19.878	-0.000	19.279
Intensity: msg A	0	6.332	12.515	0.969	2.541	6.916
Intensity: msg C	0	0.006	0.019	0.000	0.001	0.005
Intensity: msg D	0	6.060	12.028	0.900	2.407	6.630
Intensity: msg E	0	0.294	0.670	0.051	0.138	0.309
Intensity: msg F	0	0.017	0.161	0.000	0.001	0.006
Intensity: msg P	0	0.067	0.247	0.009	0.026	0.061
Intensity: msg U	0	0.523	1.187	0.072	0.193	0.518
Intensity: msg X	0	0.103	0.571	0.000	0.003	0.018
MPID orders	692	0.146	0.136	0.046	0.107	0.203
MPID shares	25,030	0.251	0.199	0.072	0.217	0.391
MPID \$ volume	25,030	0.664	0.298	0.421	0.765	0.933
MPID presence	25,051	0.056	0.086	0.010	0.029	0.066
Depth Imb.	305,161	0.034	15.134	-2.558	0.000	2.517
Depth Imb. (Wght.)	305,161	-0.001	0.289	-0.168	-0.001	0.166
HFT presence	0	0.042	0.125	0.004	0.010	0.032
Flash Crash	0	$9.34 \cdot 10^{-4}$	$3.05 \cdot 10^{-2}$	0.000	0.000	0.000
Roll	402,352	0.144	0.503	0.017	0.047	0.123
Amihud	181,301	$6.49 \cdot 10^{-8}$	$1.73 \cdot 10^{-6}$	$2.16 \cdot 10^{-9}$	$7.47 \cdot 10^{-9}$	$2.48 \cdot 10^{-8}$
Variance Ratio	839,517	0.862	238.983	0.322	0.537	0.883
VPIN	328,599	0.545	0.251	0.333	0.512	0.744

Table 3.3. Descriptive statistics of the microstructure features in 2019, dollar volume bars sampling.

The table presents mean, standard deviation, 25th, 50th and 75th percentile for the market microstructure measures calculated from the dollar volume bars sampled data in 2019. The total number of observations is 15,862,849. The number of missing observations for each feature is also provided in the table. For the detailed description of the features, refer to the section [3.4.2](#).

Measure	Missing	Mean	StDev	25th	50th	75th
Price	0	114.195	209.173	41.080	67.910	129.510
Return (1)	86,011	$2.26 \cdot 10^{-6}$	$1.09 \cdot 10^{-3}$	$-3.84 \cdot 10^{-4}$	0.000	$3.89 \cdot 10^{-4}$
Return (5)	414,563	$1.09 \cdot 10^{-5}$	$2.37 \cdot 10^{-3}$	$-9.03 \cdot 10^{-4}$	0.000	$9.35 \cdot 10^{-4}$
Depth	184,679	3.241	11.412	0.355	0.813	2.018
Up Resistance	163,882	1.633	6.445	0.167	0.393	0.996
Down Resistance	163,678	1.606	6.368	0.168	0.395	0.982
Realized Volatility	139,494	$2.50 \cdot 10^{-3}$	$2.10 \cdot 10^{-3}$	$1.30 \cdot 10^{-3}$	$1.98 \cdot 10^{-3}$	$3.05 \cdot 10^{-3}$
Quoted Spread	139,494	6.305	9.171	2.182	3.818	6.969
Realized Spread	158,217	-0.563	30.398	-13.958	-0.000	12.976
Intensity: msg A	139,494	19.295	884.480	1.173	2.938	7.093
Intensity: msg C	139,494	0.523	122.456	0.000	0.001	0.006
Intensity: msg D	139,494	13.202	592.975	1.082	2.758	6.734
Intensity: msg E	139,494	33.369	7,722.438	0.083	0.184	0.395
Intensity: msg F	139,494	0.075	19.048	0.000	0.000	0.004
Intensity: msg P	139,494	8.240	2,625.163	0.010	0.025	0.059
Intensity: msg U	139,494	2.647	232.615	0.172	0.419	0.928
Intensity: msg X	139,494	0.244	31.115	0.000	0.005	0.045
MPID orders	139,494	0.163	0.139	0.061	0.124	0.224
MPID shares	188,395	0.281	0.197	0.102	0.265	0.425
MPID \$ volume	188,395	0.722	0.269	0.537	0.836	0.944
MPID presence	188,395	0.045	0.061	0.010	0.025	0.055
Depth Imb.	154,161	-0.229	65.369	-2.901	-0.007	2.805
Depth Imb. (Wght.)	154,161	0.000	0.277	-0.175	0.000	0.175
HFT presence	139,494	0.043	0.241	0.005	0.012	0.033
Flash Crash	0	$1.67 \cdot 10^{-3}$	$4.08 \cdot 10^{-2}$	0.000	0.000	0.000
Roll	292,888	0.085	0.246	0.016	0.036	0.081
Amihud	139,494	$1.20 \cdot 10^{-8}$	$5.07 \cdot 10^{-8}$	$1.35 \cdot 10^{-9}$	$3.62 \cdot 10^{-9}$	$9.52 \cdot 10^{-9}$
Variance Ratio	772,587	0.905	296.944	0.335	0.564	0.939
VPIN	150,249	0.509	0.226	0.330	0.466	0.664

Table 3.4. Descriptive statistics of the microstructure features in 2020, dollar volume bars sampling.

The table presents mean, standard deviation, 25th, 50th and 75th percentile for the market microstructure measures calculated from the dollar volume bars sampled data in 2020. The total number of observations is 15,949,459. The number of missing observations for each feature is also provided in the table. For the detailed description of the features, refer to the section [3.4.2](#).

Measure	Missing	Mean	StDev	25th	50th	75th
Price	0	127.015	262.025	35.050	64.865	138.240
Return (1)	85,729	$-1.78 \cdot 10^{-6}$	$2.01 \cdot 10^{-3}$	$-5.94 \cdot 10^{-4}$	0.000	$5.93 \cdot 10^{-4}$
Return (5)	413,363	$-2.44 \cdot 10^{-6}$	$4.36 \cdot 10^{-3}$	$-1.41 \cdot 10^{-3}$	0.000	$1.41 \cdot 10^{-3}$
Depth	198,413	3.121	17.272	0.244	0.578	1.462
Up Resistance	183,973	1.552	9.599	0.115	0.280	0.713
Down Resistance	182,902	1.572	10.199	0.115	0.280	0.712
Realized Volatility	166,238	$4.32 \cdot 10^{-3}$	$4.40 \cdot 10^{-3}$	$1.96 \cdot 10^{-3}$	$3.17 \cdot 10^{-3}$	$5.25 \cdot 10^{-3}$
Quoted Spread	166,238	9.159	21.477	2.573	4.949	10.137
Realized Spread	180,159	-0.734	53.668	-21.399	-0.000	20.136
Intensity: msg A	166,238	18.790	763.685	1.455	3.741	9.501
Intensity: msg C	166,238	9.791	5,279.774	0.000	0.001	0.008
Intensity: msg D	166,238	13.930	448.271	1.349	3.522	8.998
Intensity: msg E	166,238	24.509	5,172.335	0.097	0.221	0.510
Intensity: msg F	166,238	0.134	32.958	0.000	0.001	0.007
Intensity: msg P	166,238	6.434	2,092.550	0.016	0.042	0.103
Intensity: msg U	166,238	2.128	218.578	0.112	0.282	0.740
Intensity: msg X	166,238	0.208	17.246	0.000	0.004	0.026
MPID orders	166,238	0.142	0.126	0.050	0.108	0.195
MPID shares	201,004	0.248	0.187	0.082	0.220	0.383
MPID \$ volume	201,004	0.670	0.290	0.429	0.771	0.930
MPID presence	201,004	0.048	0.065	0.011	0.027	0.060
Depth Imb.	220,440	-0.044	14.912	-2.778	0.000	2.700
Depth Imb. (Wght.)	220,440	-0.000	0.287	-0.174	0.000	0.173
HFT presence	166,238	0.065	0.267	0.006	0.015	0.049
Flash Crash	0	$1.75 \cdot 10^{-3}$	$4.19 \cdot 10^{-2}$	0.000	0.000	0.000
Roll	303,640	0.153	0.512	0.021	0.053	0.133
Amihud	166,238	$2.68 \cdot 10^{-8}$	$5.07 \cdot 10^{-7}$	$1.71 \cdot 10^{-9}$	$5.03 \cdot 10^{-9}$	$1.49 \cdot 10^{-8}$
Variance Ratio	781,022	1.531	1693.859	0.331	0.556	0.927
VPIN	187,593	0.490	0.230	0.307	0.443	0.645

Table 3.5. R^2_{oos} for pooled regressions in 2019 and 2020, based on time bars. The table presents the out-of-sample R^2 of the one-period forecast of such variables as the probability of a flash crash return one period ahead, the quoted spread, and the realized volatility. Lengths of the training and the testing periods for each year are presented as well.

	Return	Flash Crash	Spread	Realized Volatility
Year 2019				
Train Size, (000)	15,029.36	15,035.82	15,028.03	15,030.13
Test Size, (000)	1,670.49	1,672.16	1,668.82	1,670.05
Model	$R^2_{\text{oos}}, \%$			
Linear	0.11	6.48	58.61	91.37
LGBM(50)	0.29	49.54	65.73	95.51
LGBM(100)	0.16	49.60	66.59	95.93
LGBM(200)	0.26	49.59	67.06	96.38
NN(1)	-99.99	16.32	60.27	84.55
NN(2)	-2.04	26.40	60.09	93.28
NN(3)	-3.58	25.35	59.93	92.44
NN(4)	-0.05	23.45	58.83	-0.01
LSTM(1)	-27.13	15.80	60.56	97.25
LSTM(2)	-12.63	13.13	60.44	95.39
LSTM(3)	-9.55	13.56	61.08	97.18
LSTM(4)	-9.56	15.96	61.43	97.45
Year 2020				
Train Size, (000)	13,776.35	13,782.62	13,776.54	13,777.33
Test Size, (000)	1,531.23	1,532.72	1,531.04	1,530.25
Model	$R^2_{\text{oos}}, \%$			
Linear	0.15	6.73	13.92	97.83
LGBM(50)	0.72	47.44	17.20	95.11
LGBM(100)	0.87	47.65	17.63	95.86
LGBM(200)	1.21	47.65	17.71	96.56
NN(1)	-21.68	13.09	12.01	92.62
NN(2)	0.03	27.23	13.88	97.21
NN(3)	0.04	26.59	-11.01	96.32
NN(4)	0.00	23.53	-11.01	0.00
LSTM(1)	-9.9	10.05	13.49	96.65
LSTM(2)	-15.26	13.85	13.53	97.66
LSTM(3)	-15.38	13.97	12.62	97.71
LSTM(4)	-6.30	13.23	13.29	97.91

Table 3.6. R^2_{OOS} for pooled regressions in 2019 and 2020, based on dollar volume bars. The table presents the out-of-sample R^2 of the one-period forecast of such variables as the probability of a flash crash return one period ahead, the quoted spread, and the realized volatility. Lengths of the training and the testing periods for each year are presented as well.

	Return	Flash Crash	Spread	Realized Volatility
Year 2019				
Train Size, (000)	14,086.52	14,093.62	14,086.68	14,086.79
Test Size, (000)	1,565.77	1,568.21	1,565.61	1,565.5
Model	$R^2_{\text{OOS}}, \%$			
Linear	0.04	46.41	47.37	88.10
LGBM(50)	0.08	81.79	57.11	94.31
LGBM(100)	0.10	81.71	57.87	94.85
LGBM(200)	0.03	81.57	58.26	95.51
NN(1)	-678.88	51.19	46.50	66.49
NN(2)	-0.12	65.15	49.26	94.65
NN(3)	0.00	65.29	49.79	95.39
NN(4)	-0.01	62.36	48.16	0.00
LSTM(1)	-7.58	50.10	50.60	96.05
LSTM(2)	-1.11	63.50	50.19	91.75
LSTM(3)	-1.87	63.01	51.58	96.86
LSTM(4)	-3.29	63.24	51.48	96.65
Year 2020				
Train Size, (000)	12,891.25	12,900.10	12,890.97	12,891.29
Test Size, (000)	1,431.97	1,436.09	1,432.25	1,431.93
Model	$R^2_{\text{OOS}}, \%$			
Linear	0.07	51.68	43.14	96.52
LGBM(50)	0.42	84.77	53.17	96.91
LGBM(100)	0.54	84.68	53.86	97.25
LGBM(200)	0.60	84.63	54.34	97.50
NN(1)	-38.77	60.31	39.37	0.73
NN(2)	-0.35	69.35	44.54	94.95
NN(3)	-0.01	68.82	44.88	93.78
NN(4)	0.00	0.00	43.57	93.55
LSTM(1)	-28.39	57.56	37.76	96.30
LSTM(2)	-26.04	58.56	42.73	97.34
LSTM(3)	-20.55	58.08	46.38	97.75
LSTM(4)	-12.16	66.05	46.09	97.97

Table 3.7. Comparison of the one-period return out-of-sample prediction using Diebold-Mariano tests (time bars sampling).

The table shows pairwise Diebold-Mariano test statistics comparing the out-of-sample prediction of the pooled regressions on the return in the next time bar. The comparison is among twelve models (simple linear regression, LGBM, Neural Network, LSTM with various architectures). Positive numbers indicate the column model outperforms the row model. Bold font indicates the pairwise test difference is significant at 5% level or better.

Year 2019	LGBM (50)	LGBM (100)	LGBM (200)	NN(1)	NN(2)	NN(3)	NN(4)	LSTM (1)	LSTM (2)	LSTM (3)	LSTM (4)
Linear	2.18	0.34	0.96	-1.53	-1.29	-1.31	-3.59	-15.79	-34.26	-6.35	-16.05
LGBM(50)		-1.31	-0.28	-1.53	-1.39	-1.37	-3.57	-15.87	-33.18	-6.45	-16.19
LGBM(100)			0.73	-1.53	-1.31	-1.32	-1.42	-15.66	-33.76	-6.41	-15.95
LGBM(200)				-1.53	-1.37	-1.36	-1.98	-15.87	-30.41	-6.37	-15.67
NN (1)					1.53	1.54	1.52	1.14	1.34	1.41	1.38
NN (2)						-1.34	1.22	-75.54	-7.59	-16.02	-4.83
NN (3)							1.27	-19.96	-3.57	-4.22	-2.25
NN (4)								-16.04	-36.63	-6.39	-16.16
LSTM(1)									9.89	26.84	10.91
LSTM(2)										2.55	5.63
LSTM(3)											-0.01

Year 2020	LGBM (50)	LGBM (100)	LGBM (200)	NN(1)	NN(2)	NN(3)	NN(4)	LSTM (1)	LSTM (2)	LSTM (3)	LSTM (4)
Linear	3.43	3.79	4.55	-51.49	-5.73	-4.13	-4.71	-97.58	-62.81	-81.15	-38.87
LGBM(50)		2.80	4.68	-53.46	-3.95	-3.80	-3.99	-58.85	-65.14	-73.69	-31.01
LGBM(100)			4.87	-54.59	-4.23	-4.09	-4.27	-55.49	-67.83	-73.51	-29.49
LGBM(200)				-55.19	-4.88	-4.74	-4.90	-48.10	-65.95	-67.98	-27.07
NN (1)					50.65	50.24	50.10	28.11	15.17	14.75	33.54
NN (2)						0.93	-1.48	-94.34	-61.23	-79.91	-38.16
NN (3)							-2.35	-92.29	-60.45	-78.96	-38.79
NN (4)								-90.82	-60.36	-79.37	-37.61
LSTM(1)									-26.88	-35.37	22.85
LSTM(2)										-0.75	38.54
LSTM(3)											50.55

Table 3.8. Comparison of the mini-flash crash out-of-sample prediction using Diebold-Mariano tests (time bars sampling).

The table shows pairwise Diebold-Mariano test statistics comparing the out-of-sample prediction of the pooled regressions on the mini-flash crash occurrence probability in the next time bar. The comparison is among twelve models (simple linear regression, LGBM, Neural Network, LSTM with various architectures). Positive numbers indicate the column model outperforms the row model. Bold font indicates the pairwise test difference is significant at 5% level or better.

Year	LGBM	LGBM	LGBM	NN(1)	NN(2)	NN(3)	NN(4)	LSTM	LSTM	LSTM	LSTM
2019	(50)	(100)	(200)					(1)	(2)	(3)	(4)
Linear	20.84	20.79	20.73	14.24	17.01	16.37	16.03	13.49	13.89	14.15	12.89
LGBM(50)		0.9	0.46	-19.32	-15.78	-16.36	-16.34	-19.48	-20.23	-20.18	-19.68
LGBM(100)			-0.11	-19.26	-15.71	-16.28	-16.27	-19.42	-20.17	-20.11	-19.61
LGBM(200)				-19.18	-15.62	-16.19	-16.19	-19.33	-20.09	-20.03	-19.52
NN(1)					13.63	12.48	10.73	-1.17	-7.13	-6.49	-0.74
NN(2)						-4.32	-9.75	-14.96	-16.2	-15.8	-15.04
NN(3)							-6.68	-14.22	-15.55	-15.12	-14.22
NN(4)								-12.85	-14.61	-14.09	-12.78
LSTM(1)									-10.17	-9.53	1.26
LSTM(2)										5.79	8.73
LSTM(3)											8.24

Year	LGBM	LGBM	LGBM	NN(1)	NN(2)	NN(3)	NN(4)	LSTM	LSTM	LSTM	LSTM
2020	(50)	(100)	(200)					(1)	(2)	(3)	(4)
Linear	19.79	19.8	19.74	13.20	14.82	14.58	14.54	5.61	8.07	9.82	8.43
LGBM(50)		2.61	1.48	-19.66	-14.11	-14.45	-15.4	-20.01	-19.41	-19.92	-19.86
LGBM(100)			-0.04	-19.66	-14.12	-14.45	-15.41	-20.01	-19.41	-19.92	-19.86
LGBM(200)				-19.58	-14.02	-14.35	-15.32	-19.94	-19.33	-19.83	-19.77
NN(1)					13.87	13.51	13.94	-5.33	1.22	1.93	0.31
NN(2)						-4.76	-7.13	-14.91	-13.65	-14.37	-14.21
NN(3)							-6.13	-14.56	-13.18	-13.9	-13.78
NN(4)								-12.61	-11.54	-12.28	-14.45
LSTM(1)									7.45	9.83	5.24
LSTM(2)										0.48	-2.00
LSTM(3)											-2.44

Table 3.9. Comparison of the quoted spread out-of-sample prediction using Diebold-Mariano tests (time bars sampling).

The table shows pairwise Diebold-Mariano test statistics comparing the out-of-sample prediction of the pooled regressions on the quoted spread in the next time bar. The comparison is among twelve models (simple linear regression, LGBM, Neural Network, LSTM with various architectures). Positive numbers indicate the column model outperforms the row model. Bold font indicates the pairwise test difference is significant at 5% level or better.

Year 2019	LGBM (50)	LGBM (100)	LGBM (200)	NN(1)	NN(2)	NN(3)	NN(4)	LSTM (1)	LSTM (2)	LSTM (3)	LSTM (4)
Linear	36.25	34.87	34.14	10.59	19.94	18.10	2.91	34.09	33.86	36.09	39.33
LGBM(50)		15.57	16.30	-22.98	-27.90	-28.45	-33.34	-27.80	-28.3	-25.31	-23.77
LGBM(100)			12.10	-24.00	-27.96	-28.34	-32.42	-27.52	-27.95	-25.61	-24.35
LGBM(200)				-24.36	-27.86	-28.18	-31.87	-27.26	-27.67	-25.62	-24.46
NN (1)					-1.98	-3.65	-12.45	1.9	1.11	6.02	7.92
NN (2)						-10.13	-25.13	6.15	4.86	14.73	17.34
NN (3)							-29.84	8.54	7.51	17.28	19.33
NN (4)								24.81	25.87	32.10	32.41
LSTM(1)									-4.37	14.15	22.76
LSTM(2)										18.05	26.34
LSTM(3)											15.22

Year 2020	LGBM (50)	LGBM (100)	LGBM (200)	NN(1)	NN(2)	NN(3)	NN(4)	LSTM (1)	LSTM (2)	LSTM (3)	LSTM (4)
Linear	4.49	4.00	4.34	-43.39	-0.86	-101.75	-101.75	-17.78	-8.21	-26.80	-11.13
LGBM(50)		2.09	2.66	-7.13	-4.57	-30.12	-30.12	-4.99	-5.33	-6.27	-5.63
LGBM(100)			0.64	-6.08	-4.07	-25.32	-25.32	-4.40	-4.63	-5.41	-4.87
LGBM(200)				-6.56	-4.42	-26.64	-26.64	-4.76	-5.02	-5.85	-5.28
NN (1)					45.62	-96.83	-96.83	27.47	26.04	10.64	19.06
NN (2)						-98.22	-98.22	-7.38	-5.74	-33.94	-11.17
NN (3)							290.26	102.72	88.50	92.64	86.43
NN (4)								102.72	88.50	92.64	86.43
LSTM(1)									0.60	-15.42	-3.02
LSTM(2)										-13.13	-4.63
LSTM(3)											12.58

Table 3.10. Comparison of the realized volatility out-of-sample prediction using Diebold-Mariano tests (time bars sampling).

The table shows pairwise Diebold-Mariano test statistics comparing the out-of-sample prediction of the pooled regressions on the realized volatility in the next time bar. The comparison is among twelve models (simple linear regression, LGBM, Neural Network, LSTM with various architectures). Positive numbers indicate the column model outperforms the row model. Bold font indicates the pairwise test difference is significant at 5% level or better.

Year 2019	LGBM (50)	LGBM (100)	LGBM (200)	NN(1)	NN(2)	NN(3)	NN(4)	LSTM (1)	LSTM (2)	LSTM (3)	LSTM (4)
Linear	17.24	20.39	26.21	-40.63	16.42	7.88	-68.63	152.66	94.86	152.97	162.74
LGBM(50)		6.18	7.38	-40.05	-12.83	-18.89	-81.31	7.56	-0.55	6.87	8.10
LGBM(100)			7.33	-43.22	-15.81	-22.46	-80.02	6.17	-2.59	5.53	6.83
LGBM(200)				-48.95	-20.7	-28.15	-77.96	4.77	-5.58	4.15	5.66
NN (1)					46.57	39.79	-64.30	76.74	64.88	74.97	76.88
NN (2)						-27.15	-74.84	37.88	22.53	33.79	38.11
NN (3)							-75.52	38.31	25.64	34.9	38.41
NN (4)								73.43	72.21	72.85	73.02
LSTM(1)									-95.04	-3.27	9.87
LSTM(2)										59.7	85.66
LSTM(3)											18.83

Year 2020	LGBM (50)	LGBM (100)	LGBM (200)	NN(1)	NN(2)	NN(3)	NN(4)	LSTM (1)	LSTM (2)	LSTM (3)	LSTM (4)
Linear	-2.34	-2.18	-1.97	-83.54	-9.44	-17.47	-32.82	-1.47	-0.62	-0.96	1.74
LGBM(50)		2.79	2.77	-2.25	1.90	1.11	-48.06	3.29	2.84	2.50	2.50
LGBM(100)			2.67	-3.79	1.59	0.55	-43.65	3.28	2.77	2.37	2.37
LGBM(200)				-6.60	1.09	-0.43	-39.83	0.37	2.75	2.20	2.22
NN (1)					117.10	75.03	-31.60	5.34	22.4	57.11	132.96
NN (2)						-37.69	-33.28	-0.74	2.10	6.13	18.05
NN (3)							-33.21	0.46	6.69	19.82	28.43
NN (4)								41.13	35.59	34.06	33.25
LSTM(1)									1.76	1.56	1.65
LSTM(2)										0.27	1.05
LSTM(3)											2.19

Table 3.11. Comparison of the one-period return out-of-sample prediction using Diebold-Mariano tests (dollar volume bars sampling).

The table shows pairwise Diebold-Mariano test statistics comparing the out-of-sample prediction of the pooled regressions on the return in the next volume bar. The comparison is among twelve models (simple linear regression, LGBM, Neural Network, LSTM with various architectures). Positive numbers indicate the column model outperforms the row model. Bold font indicates the pairwise test difference is significant at 5% level or better.

Year 2019	LGBM (50)	LGBM (100)	LGBM (200)	NN(1)	NN(2)	NN(3)	NN(4)	LSTM (1)	LSTM (2)	LSTM (3)	LSTM (4)
Linear	0.69	0.70	-0.04	-1.65	-3.10	-2.60	-3.90	-3.33	-28.94	-15.95	-10.67
LGBM(50)		0.33	-0.55	-1.65	-2.50	-1.22	-1.43	-3.34	-16.60	-14.54	-10.54
LGBM(100)			-1.16	-1.65	-2.23	-1.10	-1.26	-3.35	-13.30	-13.70	-10.42
LGBM(200)				-1.65	-1.18	-0.24	-0.35	-3.32	-9.04	-11.36	-9.86
NN (1)					1.65	1.65	1.65	1.63	1.65	1.64	1.64
NN (2)						2.38	2.19	-3.26	-15.49	-13.81	-10.12
NN (3)							-0.93	-3.31	-26.46	-15.24	-10.52
NN (4)								-3.31	-28.47	-16.32	-10.75
LSTM(1)									2.84	2.61	2.15
LSTM(2)										-7.10	-7.19
LSTM(3)											-6.59

Year 2020	LGBM (50)	LGBM (100)	LGBM (200)	NN(1)	NN(2)	NN(3)	NN(4)	LSTM (1)	LSTM (2)	LSTM (3)	LSTM (4)
Linear	5.21	5.21	4.25	-14.73	-3.33	-3.77	-2.58	-55.14	-94.47	-57.95	-28.04
LGBM(50)		3.07	2.27	-14.87	-5.16	-5.51	-5.07	-54.33	-94.27	-57.66	-27.90
LGBM(100)			0.91	-14.91	-5.50	-5.53	-5.16	-55.08	-96.42	-58.67	-28.35
LGBM(200)				-14.93	-5.16	-4.58	-4.39	-54.37	-92.94	-57.35	-27.87
NN (1)					14.60	14.69	14.69	3.82	4.80	6.81	9.86
NN (2)						2.76	2.82	-53.86	-84.93	-54.73	-26.77
NN (3)							0.30	-55.75	-94.79	-58.51	-28.40
NN (4)								-56.04	-93.73	-58.76	-28.61
LSTM(1)									6.24	26.15	69.65
LSTM(2)										24.55	42.94
LSTM(3)											41.37

Table 3.12. Comparison of the mini-flash crash out-of-sample prediction using Diebold-Mariano tests (dollar volume bars sampling).

The table shows pairwise Diebold-Mariano test statistics comparing the out-of-sample prediction of the pooled regressions on the mini-flash crash occurrence probability in the next volume bar. The comparison is among twelve models (simple linear regression, LGBM, Neural Network, LSTM with various architectures). Positive numbers indicate the column model outperforms the row model. Bold font indicates the pairwise test difference is significant at 5% level or better.

Year	LGBM	LGBM	LGBM	NN(1)	NN(2)	NN(3)	NN(4)	LSTM	LSTM	LSTM	LSTM
2019	(50)	(100)	(200)					(1)	(2)	(3)	(4)
Linear	10.02	12.20	14.92	6.22	6.52	6.47	4.61	4.17	5.57	3.32	4.85
LGBM(50)		-1.45	-2.44	-19.13	-14.09	-13.67	-15.74	-19.24	-13.88	-14.44	-14.43
LGBM(100)			-2.24	-19.07	-14.00	-13.60	-15.64	-19.18	-13.82	-14.37	-14.36
LGBM(200)				-18.95	-13.83	-13.43	-15.47	-19.06	-13.68	-14.22	-14.21
NN(1)					13.37	14.17	11.48	-5.27	12.85	12.81	12.74
NN(2)						0.54	-11.46	-13.89	-3.28	-4.60	-4.35
NN(3)							-9.78	-14.39	-3.99	-5.47	-5.36
NN(4)								-12.18	2.21	1.37	2.04
LSTM(1)									13.09	13.23	13.21
LSTM(2)										-4.24	-1.51
LSTM(3)											2.30

Year	LGBM	LGBM	LGBM	NN(1)	NN(2)	NN(3)	NN(4)	LSTM	LSTM	LSTM	LSTM
2020	(50)	(100)	(200)					(1)	(2)	(3)	(4)
Linear	23.74	23.62	23.56	21.25	20.42	20.1	-25.05	19.09	20.75	18.64	18.28
LGBM(50)		-1.98	-2.26	-20.08	-16.9	-17.59	-32.15	-19.37	-19.91	-19.7	-19.17
LGBM(100)			-1.4	-19.96	-16.73	-17.42	-32.11	-19.28	-19.8	-19.61	-19.02
LGBM(200)				-19.88	-16.62	-17.31	-32.11	-19.21	-19.73	-19.54	-18.91
NN(1)					15.23	13.62	-26.17	-10.06	-9.88	-10.57	11.17
NN(2)						-3.7	-29.1	-14.29	-15.16	-14.73	-14.55
NN(3)							-29.82	-13.29	-13.92	-13.49	-9.78
NN(4)								25.21	25.62	25.15	27.75
LSTM(1)									7.69	5.06	11.71
LSTM(2)										-5.83	12.29
LSTM(3)											12.09

Table 3.13. Comparison of the quoted spread out-of-sample prediction using Diebold-Mariano tests (dollar volume bars sampling).

The table shows pairwise Diebold-Mariano test statistics comparing the out-of-sample prediction of the pooled regressions on the quoted spread in the next volume bar. The comparison is among twelve models (simple linear regression, LGBM, Neural Network, LSTM with various architectures). Positive numbers indicate the column model outperforms the row model. Bold font indicates the pairwise test difference is significant at 5% level or better.

Year 2019	LGBM (50)	LGBM (100)	LGBM (200)	NN(1)	NN(2)	NN(3)	NN(4)	LSTM (1)	LSTM (2)	LSTM (3)	LSTM (4)
Linear	51.00	46.19	43.79	-3.85	9.16	16.60	4.35	36.77	29.12	40.70	36.01
LGBM(50)		12.38	12.47	-40.90	-31.23	-34.62	-36.99	-38.44	-39.81	-34.26	-35.37
LGBM(100)			8.99	-39.17	-30.66	-32.78	-35.11	-35.03	-36.16	-31.49	-32.70
LGBM(200)				-38.19	-30.17	-31.76	-34.11	-33.43	-34.44	-30.2	-31.26
NN (1)					27.64	36.41	18.78	24.36	23.21	30.25	31.46
NN (2)						5.84	-17.34	8.45	6.22	14.94	14.59
NN (3)							-25.30	8.35	4.50	17.47	17.78
NN (4)								17.75	15.92	24.95	24.04
LSTM(1)									-10.02	22.20	15.95
LSTM(2)										28.24	23.95
LSTM(3)											-2.20

Year 2020	LGBM (50)	LGBM (100)	LGBM (200)	NN(1)	NN(2)	NN(3)	NN(4)	LSTM (1)	LSTM (2)	LSTM (3)	LSTM (4)
Linear	29.82	27.44	26.71	-17.30	7.60	11.89	2.91	-31.48	-2.72	21.55	18.20
LGBM(50)		7.40	7.69	-41.50	-27.37	-26.91	-30.15	-43.99	-33.40	-25.20	-28.00
LGBM(100)			5.94	-38.31	-25.92	-25.20	-27.99	-39.58	-30.33	-23.34	-25.61
LGBM(200)				-36.91	-25.51	-24.72	-27.27	-37.82	-29.20	-22.7	-24.76
NN (1)					65.25	55.72	50.26	-7.63	25.25	45.13	39.95
NN (2)						7.50	-17.10	-34.03	-16.33	15.90	12.36
NN (3)							-36.43	-40.22	-23.61	15.81	11.07
NN (4)								-32.48	-8.66	25.86	19.55
LSTM(1)									39.07	45.37	42.95
LSTM(2)										32.73	27.16
LSTM(3)											-5.64

Table 3.14. Comparison of the realized volatility out-of-sample prediction using Diebold-Mariano tests (dollar volume bars sampling).

The table shows pairwise Diebold-Mariano test statistics comparing the out-of-sample prediction of the pooled regressions on the realized volatility in the next volume bar. The comparison is among twelve models (simple linear regression, LGBM, Neural Network, LSTM with various architectures). Positive numbers indicate the column model outperforms the row model. Bold font indicates the pairwise test difference is significant at 5% level or better.

Year 2019	LGBM (50)	LGBM (100)	LGBM (200)	NN(1)	NN(2)	NN(3)	NN(4)	LSTM (1)	LSTM (2)	LSTM (3)	LSTM (4)
Linear	85.34	96.62	87.30	3.16	39.11	38.99	-22.22	85.34	63.44	36.70	40.44
LGBM(50)		2.09	1.93	-3.06	0.21	0.64	-60.75	2.75	-2.80	1.40	1.42
LGBM(100)			1.81	-3.13	-0.14	0.38	-53.14	3.05	-4.68	1.27	1.29
LGBM(200)				-3.22	-0.83	-0.11	-45.08	2.96	-12.10	1.08	1.10
NN (1)					3.13	3.21	-7.07	3.32	2.81	3.53	3.40
NN (2)						15.49	-30.57	1.32	-3.95	4.01	10.13
NN (3)							-30.57	0.61	-4.78	2.69	6.29
NN (4)								45.24	38.24	29.88	31.17
LSTM(1)									-12.8	0.67	0.58
LSTM(2)										5.16	6.52
LSTM(3)											-0.58

Year 2020	LGBM (50)	LGBM (100)	LGBM (200)	NN(1)	NN(2)	NN(3)	NN(4)	LSTM (1)	LSTM (2)	LSTM (3)	LSTM (4)
Linear	1.89	4.14	6.51	-1.13	-18.39	-32.52	-50.42	-3.48	26.61	40.47	49.87
LGBM(50)		8.25	8.12	-1.14	-10.20	-18.20	-18.75	-3.38	2.17	4.07	5.21
LGBM(100)			6.77	-1.14	-13.61	-23.17	-24.06	-6.13	0.50	2.75	4.04
LGBM(200)				-1.15	-17.42	-29.29	-30.87	-9.08	-1.11	1.59	3.09
NN (1)					1.12	1.10	1.10	1.13	1.14	1.15	1.15
NN (2)						-28.08	-19.01	14.74	25.08	27.29	30.98
NN (3)							-4.77	33.98	39.40	39.71	44.96
NN (4)								50.25	65.51	61.48	74.89
LSTM(1)									15.22	20.51	25.11
LSTM(2)										17.43	31.22
LSTM(3)											14.88

Figure 3.1. Feed-forward Neural Network.

The figure presents a diagrams of a simple neural networks with two hidden layers. Green circles denote the input layer, blue circles denote the hidden layers, and red circle denotes an output layer. Also, bias nodes all hidden layers and the output layer are represented by the yellow circles. Each arrow is associated with a weight parameter. In the network with a hidden layer, a nonlinear activation function transforms the inputs before passing them on to the output.

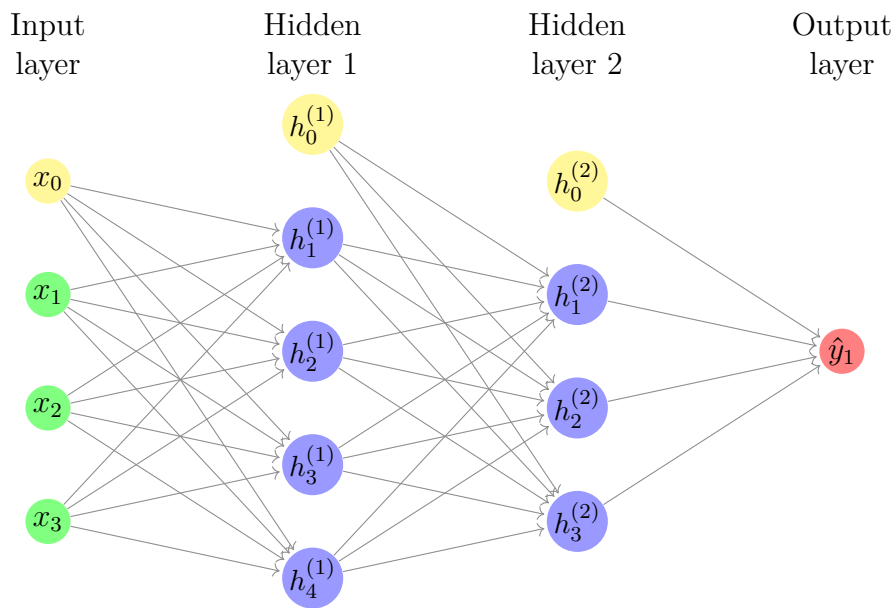


Figure 3.2. LSTM Neural Network.

The figure shows the architecture of a typical LSTM block that contains the gates σ_1 through σ_3 , the input signal $x^{(t)}$, the output $y^{(t)}$, and the activation functions g_1 and g_2 that are usually realized in the form of hyperbolic tangent. The output of the block is then connected to the next block's input. The input gate is represented by the sigmoid activation function σ_1 . This gate combines the current input and the output the LSTM unit in the last iteration. The forget gate σ_2 , determines which information should be removed from the previous cell states of the LSTM model. The output gate σ_3 , combines the current input and the output of the LSTM unit in order to calculate the current output.

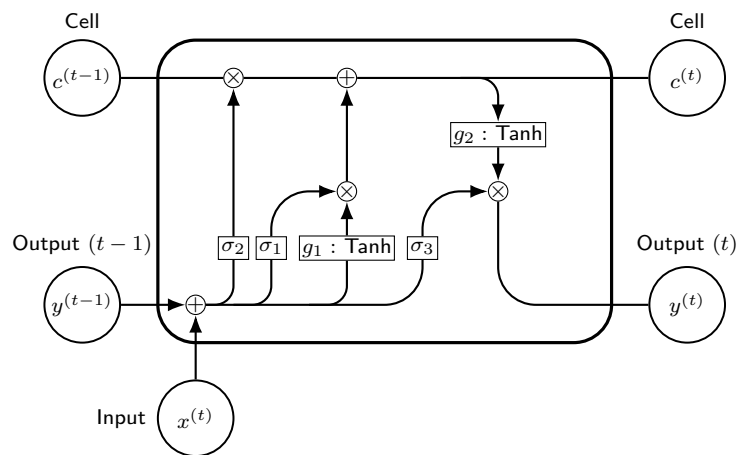


Figure 3.3. Feature importance for one-period return forecasts by sampling and year.

The Figure presents the relative feature importance for the out-of-sample forecast of one-period return. The calculation of the importance is based on a reduction in the out-of-sample R^2 when the feature was removed. Features importance is calculated for the best model and in each case is normalized to sum to 100%.

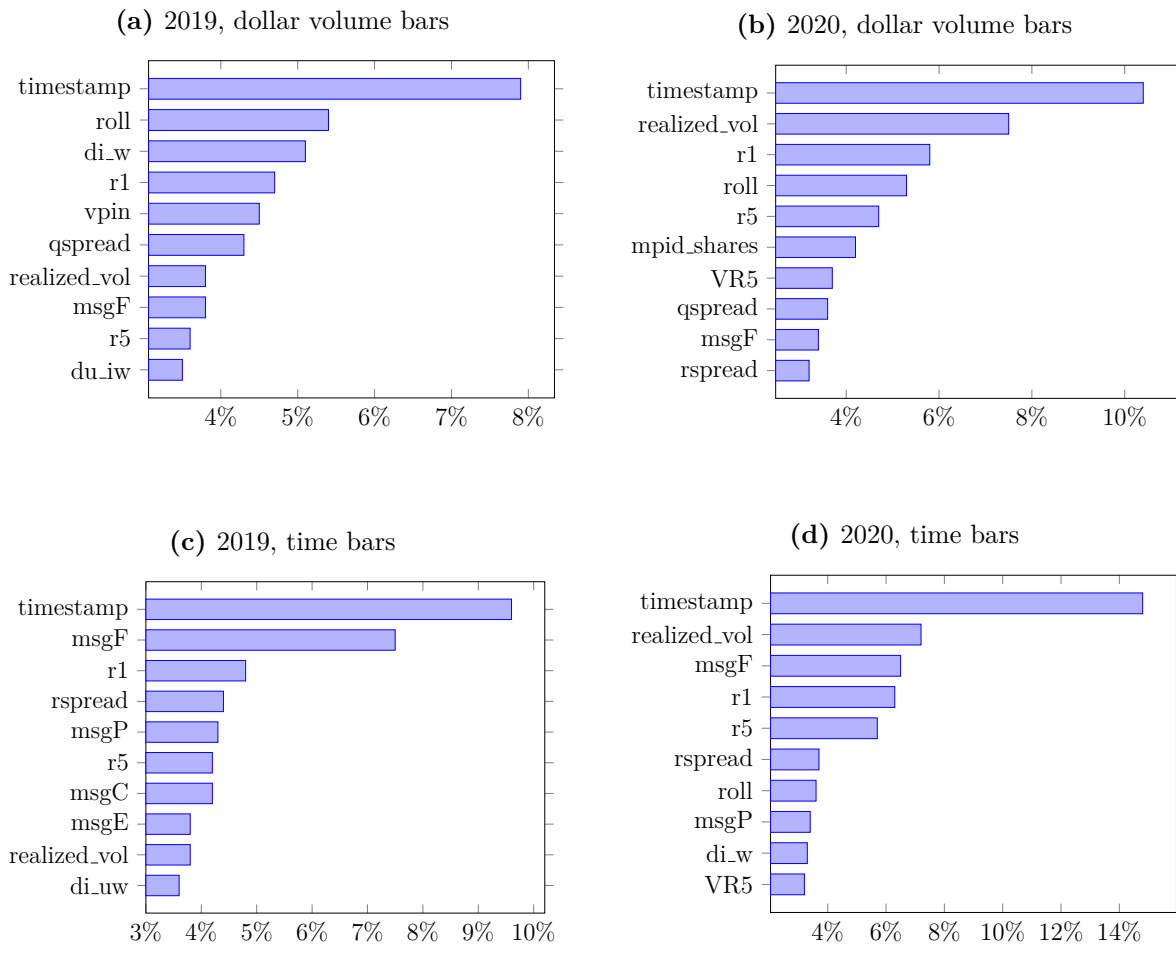


Figure 3.4. Feature importance for mini-flash crash occurrence forecasts by sampling and year.

The Figure presents the relative feature importance for the out-of-sample forecast of the mini-flash crash occurrence. The calculation of the importance is based on a reduction in the out-of-sample R^2 when the feature was removed. Features importance is calculated for the best model and in each case is normalized to sum to 100%.

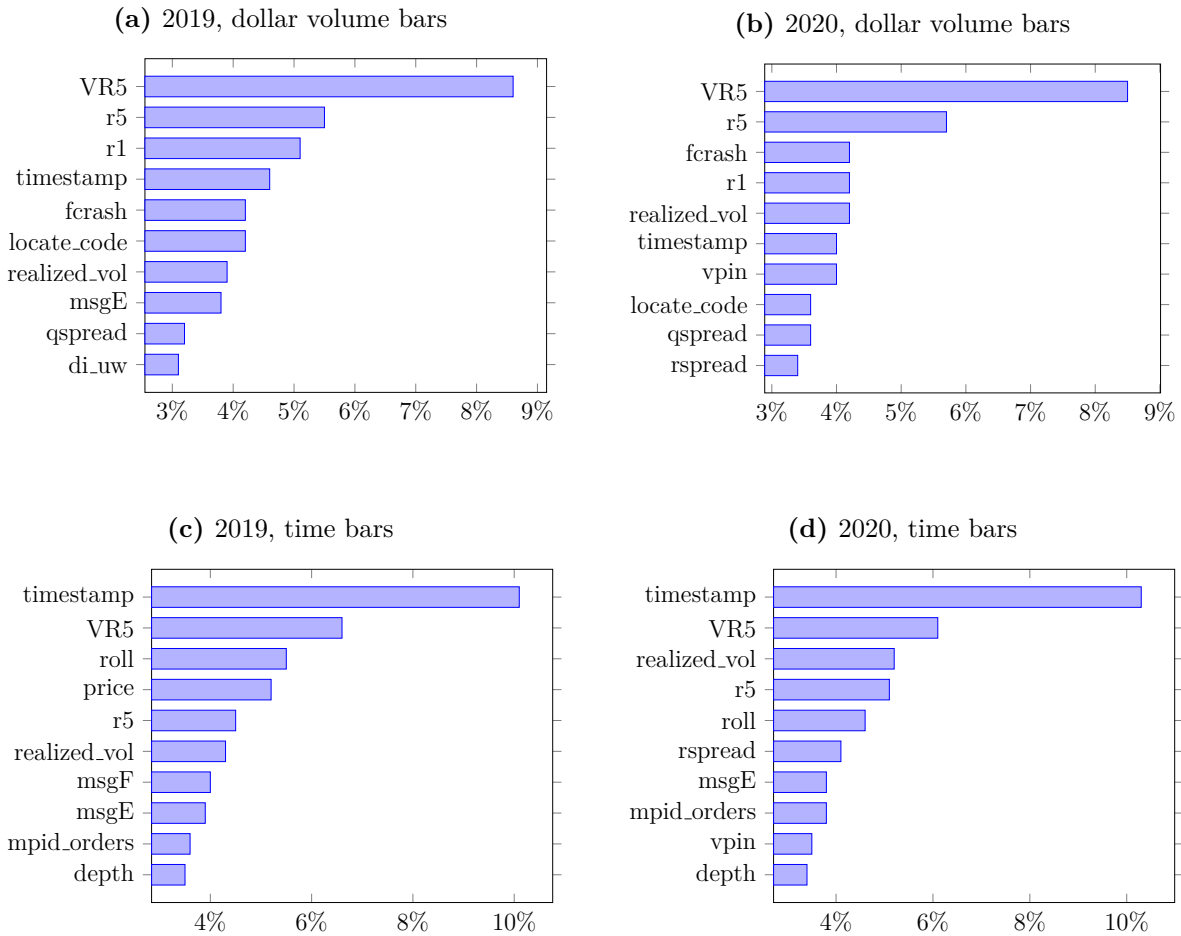


Figure 3.5. Feature importance for quoted spread forecasts by sampling and year.

The Figure presents the relative feature importance for the out-of-sample forecast of the quoted spread. The calculation of the importance is based on a reduction in the out-of-sample R^2 when the feature was removed. Features importance is calculated for the best model and in each case is normalized to sum to 100%.

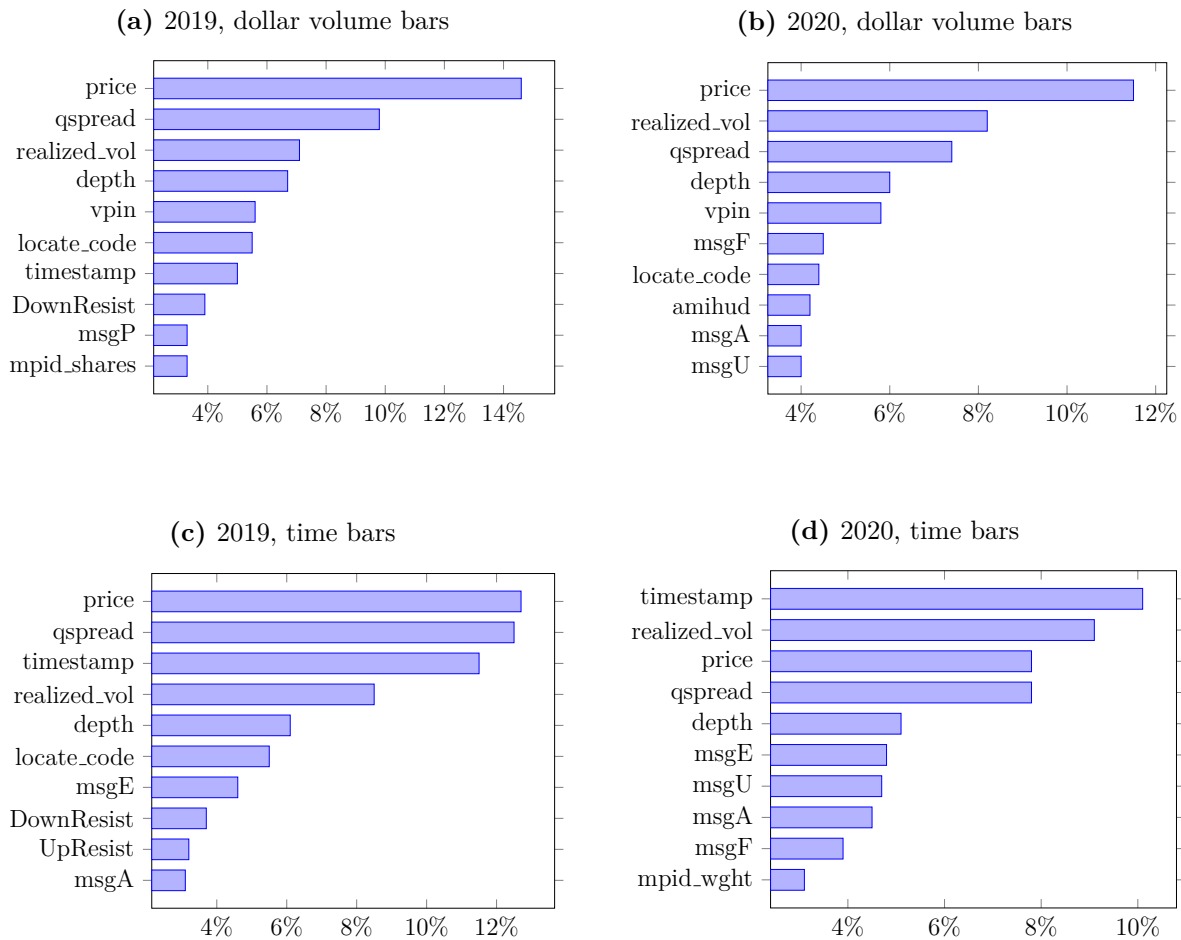
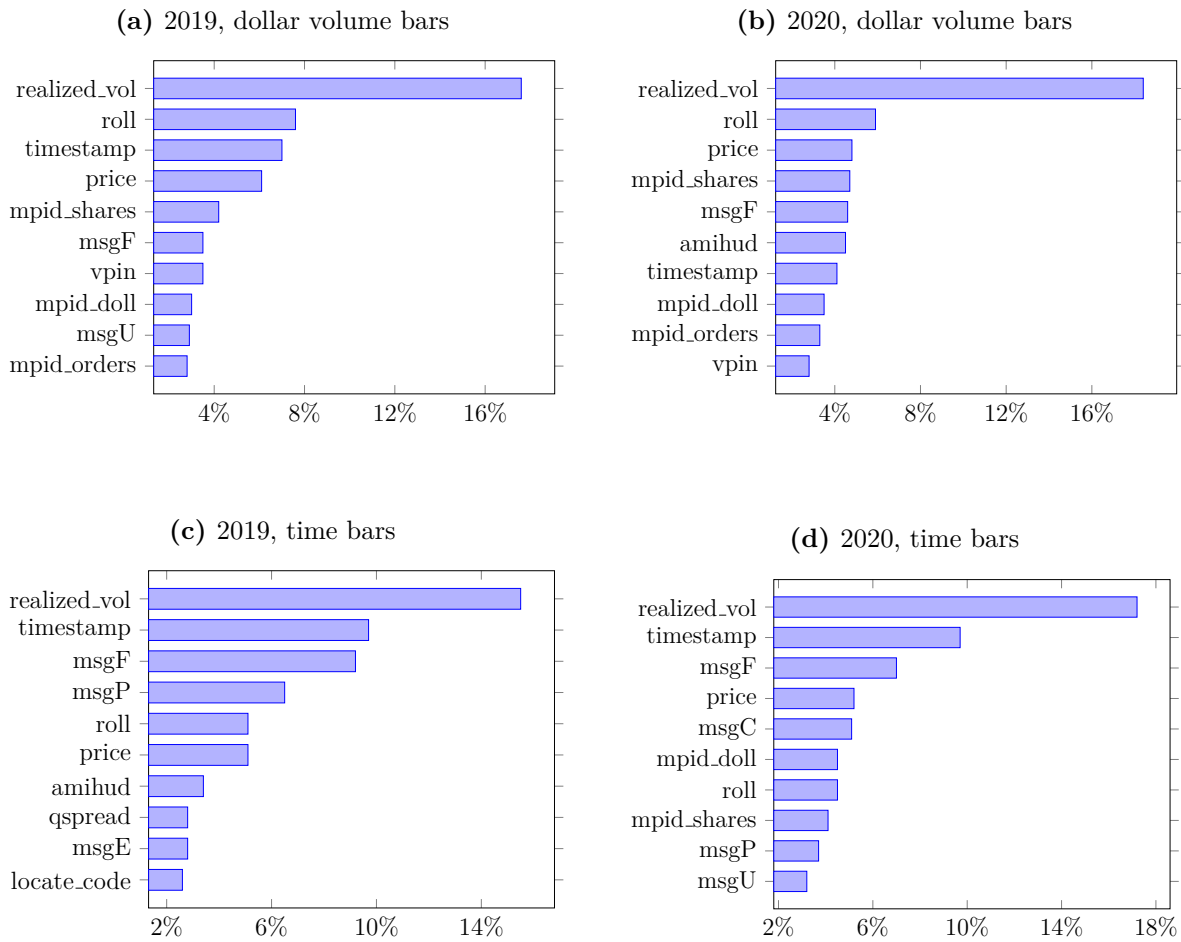


Figure 3.6. Feature importance for realized volatility forecasts by sampling and year.

The Figure presents the relative feature importance for the out-of-sample forecast of the realized volatility. The calculation of the importance is based on a reduction in the out-of-sample R^2 when the feature was removed. Features importance is calculated for the best model and in each case is normalized to sum to 100%.



3.A Appendix

3.A.1 Nasdaq order book messages

Table A.1. Nasdaq order book messages.

The table shows the types of messages disseminated by the exchange and their description provided by Nasdaq.

Message Type	Name	Description
A	Add Order – No MPID Attribution	A new order has been accepted by the Nasdaq system and was added to the displayable book. No market participant identification (MPID) was provided.
F	Add Order with MPID Attribution	A new order with market participant identifier was added to the book.
E	Order Executed Message	An order on the book is executed in whole or in part. The multiple Order Executed Messages on the same order are cumulative.
C	Order Executed with Price Message	An order on the book is executed in whole or in part at a price different from the initial display price. It is possible to receive multiple Order Executed and Order Executed With Price messages for the same order if that order is executed in several parts.
X	Order Cancel Message	An order on the book is modified as a result of a partial cancellation.
D	Order Delete Message	An order on the book is being cancelled. All remaining shares are no longer accessible so the order must be removed from the book.
U	Order Replace Message	An order on the book has been canceled and replaced. All remaining shares from the original order are no longer accessible, and must be removed. The new order details are provided for the replacement.
P	Trade Message	A non-displayable (hidden) order was executed in whole or in part. It is possible to receive multiple Trade Messages for the same order if that order is executed in several parts.

3.A.2 Details on Microstructure Measures employed

Short Name	Full Name and Description
price	Price. The last price from the trade within a bar (time or dollar volume).
qspread	Quoted spread. The difference, in basis points, between the best bid and the best ask price in the limit order book.
realized_vol	Realized volatility within a bar.
rsread	One bar realized spread.
locate_code	Locate Code. A unique reference code provided by Nasdaq to each stock.
timestamp	Timestamp. Time (in nanoseconds) passed from the midnight.
DownResist	Down Resistance. The amount in dollars required to move the price down by 50 basis points.
UpResist	Up Resistance. The amount in dollars required to move the price up by 50 basis points.
msgA	The intensity of A messages arrival to the exchange.
msgC	The intensity of C messages arrival to the exchange.
msgE	The intensity of E messages arrival to the exchange.
msgP	The intensity of P messages arrival to the exchange.
msgU	The intensity of U messages arrival to the exchange.
mpid_shares	The fraction of volume added to the limit order book by identified market participants within a bar.
mpid_doll	The fraction of dollar volume added to the limit order book by identified market participants within a bar.
mpid_wght	Weighted presence of the identified market participants in the limit order book.
mpid_orders	The fraction of limit orders submitted by identified market participants within a bar.
msgF	The intensity of F messages arrival to the exchange.
amihud	Amihud measure of illiquidity withing a bar.
roll	Roll measure within a bar.
r1	One-period return.

<code>r5</code>	Five-periods return.
<code>VR5</code>	Variance ratio.
<code>vpin</code>	Volume synchronized probability of the informed trading within a bar.
<code>depth</code>	Depth. Available market depth in dollars 50 basis points around the midquote.
<code>di_w</code>	Weighted depth imbalance. Weighted by the distance to the midquote difference of volumes available on bid and ask sides of the limit order book.
<code>di_uw</code>	unweighted depth imbalance. Simple difference between the dollar volume on ask and bid sides of the book.
<code>hft</code>	HFT presence within a bar.
<code>fcrash</code>	Mini-flash crash indicator.

General Conclusion

The three chapters of the present thesis address some of the crucial economic questions related to financial markets stability in the context of market microstructure. The first chapter provides novel evidence on market stability and liquidity provision due to the implementation of a non-displayed (dark) Midpoint Extended Life Order (M-ELO). M-ELO is a dark order that cannot interact with lit (visible) orders. It also possesses the speed bump effect due to the holding period prior to the execution. We use high-frequency order book message data from the Nasdaq exchange for the three years of M-ELO existence. The rule change applied on May 11, 2020, makes it possible to disentangle the dark and the speed bump impacts of M-ELO orders on market stability and liquidity.

For the period from January 2018 to October 2020, the degree of M-ELO activity is associated with a lower frequency of mini-flash crashes for Nasdaq traded securities. Results from panel regressions suggest the presence of significant effects of the M-ELO trading on crash returns, volatility, and trading activity. Higher relative volumes traded via M-ELO are associated with less turbulent crashes, which is more desirable for long-term investors. The effect of the M-ELO on the quality of the liquidity provision is mixed. We document the increase of both quoted spread and depth close to the midquote due to the M-ELO trading. At the same time, increased order imbalance may undermine the positive effect of the improved market depth.

The analysis shows that trading activity in M-ELO impacts market stability and liquidity mainly due to the speed bump effect. The reduction in the M-ELO's holding period by 98% decreases the influence of M-ELO on the market by 80% on average. The robustness of the results to different specifications of the model strengthens the conclusion that only about 20% of the M-ELO market stabilizing effect comes from its dark properties and 80% from the speed bump properties.

As M-ELO volumes are relatively small, we are cautious about extrapolating the results

of this analysis. The main goal of our research is to identify the effects of M-ELO on market stability during recent years. Our study delivers an important insight for market participants, policymakers, and researchers. The trade-off between the execution speed and order transparency is capable of impacting the general stability of financial markets.

The second chapter analyzes the impact of dark pool operating alongside a transparent limit order exchange on market quality and price discovery. In the four-period model with asymmetric information about the final pay-off of the risky asset, fully rational, risk-neutral traders can choose order type and venue to send their orders to. I find that the order flow migrates to the dark pool in the last periods, as the limit order book fills.

When orders are directed to the opaque venue, the process of price discovery slows down. However, in relative terms, the effect is close to negligible and may vary in nature (benefiting price discovery) depending on the immediacy demands of investors. As limit orders provide a substitute for dark orders, venue competition plays a significant role in the impact of dark trading on price discovery and market quality. When information asymmetry increases the order migration forms a U-shape and the uninformed investors are quickly crowded out from the dark pool.

The model suggests that a dark pool affects market quality mildly. The execution quality of market orders measured by the average quoted bid-ask spread and the average depth is stable in the wide range of the dark pool availability. Also, the quality of limit orders, measured by the average fill rate, slightly increases with the more available dark venue. Overall, the total welfare remains leveled for various degrees of dark pool availability.

The developed model is flexible and allows for a wide range of analyses that interest policymakers and regulators. However, future work can extend the model by addressing the following caveats. Firstly, since traders enter the model only once, I do not allow for price manipulation. If a trader can submit an order several times during the trading day, she might choose to trade first on the lit exchange to move the market and then execute later in the dark pool at a more comfortable price.

Secondly, one might be interested in endogenizing private information acquisition and dark pool participation. By introducing a cost of obtaining perfect information about the final pay-off of the asset and a cost of accessing the dark pool it will be possible to identify an equilibrium level of dark trading.

Finally, more types of orders can be added to the model to bring it closer to the current state of financial markets where the competition between transparent and dark venues is more subtle. Some exchanges allow traders to submit undisplayed orders that offer a price improvement and execute against the visible order flow. The introduction of such order types to the present model would be interesting future research.

In the third chapter, several machine learning approaches with different architectures were tested against each other and against the simple linear regression to find out which one provides more accurate forecasts of the market microstructure variables. Different targets for the forecasting were used that describe proxies for market volatility, stability, and liquidity. Out of all tested models, the Light Gradient Boosted Machine (LGBM) model demonstrated superior performance most of the time for all targets. Long Short Term Memory (LSTM) neural network proved to be the best for the autocorrelated targets such as realized volatility, while a simple feed-forward neural network showed mixed results due to frequent overfitting training data.

The advantages of dollar volume bars sampling against time bars were also examined. The results show that dollar volume bars might work better for the prediction of variables for which the trading volume is an important component as the mini-flash crash example suggests. Moreover, dollar volume bars appear to be a more stable way of data sampling for the prediction of microstructure measures during times of market turbulence. The prediction accuracy of the quoted spread reduced dramatically in 2020 for the time bars but stayed stable for volume bars.

The question, of whether it is easier to predict the microstructure variables during crisis times is currently answered in the negative. The out-of-sample R^2 for the forecasts does not suggest worse performance in 2020 compared to 2019 for almost all models. For the quoted spread, the reduction in the prediction accuracy may be mitigated by using dollar volume bars sampling.

In conclusion, the above chapters describe both the theoretical and empirical research work that has been done on dark trading and high frequency trading impact on financial markets. I hope, the analysis and the conclusions derived from the models employed are shedding more light on the important questions of market stability and liquidity, and are contributing to making financial markets safer.