This thesis studies how recent technological advancement has shaped the structure of financial markets. Three chapters are collated, each focusing on a particular aspect of the modern financial market: 1) a detailed analysis of the Flash Crash on May 6, 2010; 2) a theory on how the interaction within the market intermediary sector affects market quality; and 3) a new modeling framework for limit order market.

The author holds a bachelor degree from the University of Hong Kong, an M.Phil. degree from Tinbergen Institute, and Ph.D. degree from VU University Amsterdam. Financial support from Duisenberg School of Finance during the author’s graduate study is gratefully acknowledged.
FRictions in Modern Financial Markets and the Implications for Market Quality

A Thesis

by

Bart Zhou Yueshen*

VU University Amsterdam,
Tinbergen Institute, and
Duisenberg School of Finance

* Officially known as Zhou (family), Yueshen (given); 周越申 in Chinese characters.
ISBN: 978 90 361 0397 8

Cover design: Crasborn Graphic Designers bno, Valkenburg a.d. Geul

This book is no. 588 in the Tinbergen Institute Research Series, established through cooperation between Thela Thesis and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.
VRIJE UNIVERSITEIT

FRICTIONS IN MODERN FINANCIAL MARKETS
AND THE IMPLICATIONS FOR MARKET QUALITY

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. F.A. van der Duyn Schootuten,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Economische Wetenschappen en Bedrijfskunde
op dinsdag 24 juni 2014 om 15.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Yueshen Zhou

geboren te Shanghai, China
promotor: prof.dr. A.J. Menkveld
other committee members: prof.dr. T. Foucault
prof.dr. P.A. Gautier
dr. V. van Kervel
dr. H. Zhu
To the part of me
    that is inspired by and devoted to
    science:

    A giant though I might not become, a dwarf would still have its measurable height.
    Proudly, I shoulder up the knowledge base of all human beings by my inch—well, maybe a few feet.
# Contents

Dedication ........................ v
Preface ................................ xi

1 Road map .......................... 1

2 The Flash Crash ................. 5
   2.1 Introduction .................. 6
   2.2 Data .......................... 10
   2.3 Narrative of the Flash Crash . 11
      2.3.1 Exploratory analysis of the large seller’s trades .... 11
      2.3.2 High-frequency analysis of price changes and net order flows ... 13
         2.3.2.1 Price cointegration ........................... 15
         2.3.2.2 VAR/VECM analysis of price and net flow ........ 17
      2.3.3 Characterization of the large seller’s algorithm ........ 26
         2.3.3.1 The large seller’s size in the market .......... 26
         2.3.3.2 Predictors of large seller aggressiveness ..... 28
   2.4 Price for immediacy .......... 32
   2.5 Conclusion ................... 37
   2.6 Appendix ..................... 38
      2.6.A Data integrity ........... 38
      2.6.B Breakpoints estimation .... 43

3 Middlemen Interaction .......... 45
   3.1 Introduction ................ 46
   3.2 Model primitives .......... 50
   3.3 Equilibrium ................ 53
      3.3.1 M in the interim period .... 53
      3.3.2 S in the ex post period ... 54
         3.3.2.1 Fully revealing .......... 55
         3.3.2.2 Partial learning ....... 55
CONTENTS

3.3.2.3 No learning ............................................ 59
3.3.3 S in the ex ante period .................................. 61
3.4 Welfare ......................................................... 62
3.5 Discussion ..................................................... 64
3.6 Conclusion ..................................................... 66
3.7 Appendix ......................................................... 67
  3.7.A Notation summary ........................................ 67
  3.7.B Proofs ....................................................... 68
  3.7.C Probability of an MM trade conditional on an interim trading price .... 72
  3.7.D Solution to S’s ex ante optimization problem ..................... 73

4 Queuing Uncertainty ................................................. 75
  4.1 Introduction ................................................. 76
  4.2 Related literature ........................................... 80
  4.3 Baseline model .............................................. 82
    4.3.1 Model setup ........................................... 82
    4.3.2 Equilibrium ............................................ 84
    4.3.3 Examples ............................................... 87
  4.4 Book depth dynamics ........................................ 87
    4.4.1 Model setup ........................................... 89
    4.4.2 Equilibrium ............................................ 91
    4.4.3 Latency and order book depth dynamics ....................... 95
    4.4.4 Policy discussion ...................................... 98
  4.5 Endogenous market order and welfare ......................... 99
    4.5.1 Equilibrium ............................................ 100
    4.5.2 Liquidity provision and welfare ........................ 101
    4.5.3 Market design ......................................... 104
    4.5.4 Numerical example .................................... 106
  4.6 Deep in the book ........................................... 108
    4.6.1 The investor’s order size ................................ 108
    4.6.2 Depths in the order book ................................ 109
    4.6.3 Alternative setups ..................................... 112
  4.7 Conclusion ..................................................... 113
  4.8 Appendix ......................................................... 115
    4.8.A Speed and queuing ....................................... 115
    4.8.B Notation summary ....................................... 116
    4.8.C Proofs ..................................................... 118

Bibliography .......................................................... 129
<table>
<thead>
<tr>
<th>CONTENTS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Summary in Dutch</td>
<td>139</td>
</tr>
<tr>
<td>Acknowledgment</td>
<td>143</td>
</tr>
</tbody>
</table>
Now that I am writing this preface, many of the chances to challenge the destiny have gone. Like a maturing option contract, one’s value reduces as the possibility to go wild diminishes. (The only thing that still matters is whether it is in- or out-of-the-money.) Yet, history is not easily forgotten. Few people, as far as my life has encountered, live as free and easy as a Markov process that permits little regret. Sadly and happily—out of my oxymoron nature—I admit that I am of no exception. Hence, in retrospect, there rises a perhaps meaningful question to ruminate: Was I throughout so determined to pursue an academic path in finance?

The fact, as certified by the possibly still amenable memory (or illusions) in my brain, is that there once were many alternatives: To mathematics, literature, or even philosophy I could have turned. Fortunately or maybe not, most people are (perhaps too) susceptible to what are given; and so am I—the main theme in my education so happened to be economics and finance, which I quite happily followed and pursued for years, and now eventually am ready to push whose frontier forward.

The hindsight: The particular path that I have been through appears to have certain “nice” equilibrium feature—like an attractive fixed point\(^1\)—in the sense that the more intimately I embrace myself with economic thinking, the stronger I feel its attraction: the beautiful, mature scientific approach; the insights that I derive from textbooks and apply to the real life; and—perhaps the most snobbish reason—the cliquish satisfaction I enjoy when showing off my ability to reason in economics.

Surely, it is a matter of comparative advantage\(^2\). If randomness is the nature of the universe, I have little doubt that I have been endowed with an attribute set overlapping sufficiently with the set of so-called “economists”, such that my comparative advantage happens to fall in the field of economics, financial economics in particular. Hence, along

---

\(^1\) An attractive fixed point can be understood as an iteration through which the process reinforces itself. For example, if one who studies economics keeps studying the subject indefinitely, then “to study economics” is an attractive fixed point. This example applies to me. For those who it does not apply to, please consider getting further closer to economics, as the initiation neighborhood is usually tight, i.e. the attracting property is local only.

\(^2\) One who has a comparative advantage in some activity incurs a lower (marginal) opportunity cost, than someone else, if s/he pursues that activity.
The years of study, a self-fulfilling prophet realizes: “Thou shalt become an economist, and thus wilt thou”.

The field is large and, sometimes, unfathomable. My limited talent concludes it impossible for me to devote fully into all subfields in economics, and thus I humbly pick my specialization to study financial securities markets and the trading process. The decision was made about four years ago, with the dramatic May 6 Flash Crash thundering in the background (see chapter 2 for an empirical analysis), probably serving as an implicit trigger. The self-fulfilling mechanism (or perhaps the susceptibility in me), again, worked like a charm: The further I went into the field, the more fascinating I felt about it.

There are at least three good reasons for my happy drowning into the study of financial securities markets: First, the direct focus on trading process necessarily brings the attention to the very details of the real world: Where do people really “meet” and trade? How are trading platforms organized? What are the rules of trading games? Based on my dive into the ocean of the literature—certainly shallow my dive is and thus very confined my eye sight must be—these aspects are largely underemphasized. The appealing handiness of Walrasian clearing, Nash bargaining, perfect competition, and the alike abstracts from the binding realities, which could be and possibly are of first-order concerns, as alerted by the Flash Crash (and the many mini flash crashes), by the still ongoing debate on high-frequency trading, and by the capricious regulations (e.g. the recent trading tax implemented in France). The study of financial markets enjoys the privilege to address reality in such a direct fashion and this characteristic enables the research to straightforwardly speak to both practitioners and regulators to prove its usefulness.

In collaboration of the first point above, this second point is the other side of the coin. Financial securities markets have always remained vigorous and its development consistently sparks researches in the vanguard. As the latest technology is incorporated into trading, participants adapt to the evolving environment and the policy-makers move to patch the regulations to cope with the expected and/or unintended consequences. A large number of news reports center around such development in the market. It is thus advantageous to find novel, pertinent research topics in this field; new ideas easily blossom. The three essays collated in this thesis all benefit from such feedback from the real-world market observations: They all identify frictions from recent development in financial markets and in trading technology and then make further contribution to the understanding of market quality. Fortunately, thanks to such inspiration from the development of financial securities trading, all

---

1 Lacking a better word, I choose to use “unfathomable” here, with the hope to sketch the Chinese expression of “水很深”.

2 The relevance of market structure is a particularly important aspect in modern financial securities transaction, the process of which is populated by sophisticated “investors” of various kinds. As O’Hara (2014) points out, such sophisticated investors “maximizes against market design” and hence “only microstructure matters” in modern times.
three papers have earned certain recognition which I am honored to accolade.\footnote{The mere honor should be shared with many of those who helped me in accomplishing these papers. As an indication of my sincere gratefulness, each of the three chapters in this thesis begins with a paragraph that is dedicated to these people. A more comprehensive acknowledgment, yet still incomplete, is found in the back matter of this thesis.}

Third, the abundance of trading data dwarfs other fields in finance. Given the current speed of trading (at latency of nanoseconds), by the time one finishes reading this line, tons of market activity will have already occurred and, most importantly, been recorded. The astronomical amount of trading data on the one hand calls for cutting edge econometrics to process, hence pushing forward the research on the empirical side. On the other hand, the overwhelming waves of data often call for help from theory to provide anchors in solid economic rationale and to guide research directions. To this extent, financial securities trading possesses an ideal research feature that the development of both theory and empirics propel each other. This feature is probably best exemplified by my two research projects, now adapted into chapters 2 (empirical) and 3 (theory), both of which share the same root in some in-depth consideration and reflection on the Flash Crash.

As this thesis is to be submitted to fulfill the requirement for a Ph.D. degree, namely, “Doctor of Philosophy”, I would like to drone on a bit further to elaborate my own philosophical thinking on the research and study of financial securities trading and the markets. This field is sometimes known as “market microstructure”, as it often zooms in on the very details of the market: each single quote of price, each single trade, and each single participant, all of which are the atomic building blocks of the finance world. The market microstructure study is thus like a set of microscope that enables a new perspective of the financial market.

Micro though the name is, the importance of such a tool and the perspective furthered by it should not be underestimated. For one thing, it is these single quotes, single trades, and single participants that aggregate to the overview of the finance industry. Think about the summary statistics that are often used: The closing price, i.e. the very last single trade of the day, of an equity is used to mark the company’s capitalization to the market; a time series of such closing prices are employed to estimate asset pricing models, which further affects the investment decisions of both households and institutions; (co)variances of the returns measure the risk exposure of the asset; and a stock market index can sometimes reflect the health of the economy. These granted uses of summary statistics all reside on these each single, micro activity in the market.

I often ask myself: Without a thorough understanding of the activity observed under the microscope, how comfortable do I feel about reading and using the overview from the summary statistics? Though my own answer is “not that comfortable”, it is probable that everyone will have a different opinion on this issue. It is indeed possible, under certain strong assumptions, that in aggregate the frictions at the micro level will average out, biasing
not the summary statistics. So, does microstructure really matter?

I have the following argument: Consider the contrast between the beautiful order established by Newtonian physics versus the chaotic nature analyzed by the subatomic physics. Given that the former explains the daily life largely successfully, one could doubt the value of the later. Yet it is a legitimate scientific quest and I doubt whether any serious scientists would question the value of such quest. Likewise, so is the study of market microstructure: a legitimate and valuable scientific quest for a thorough understanding of the very building blocks of the modern finance world.

If I may refer to the microscope as a bottom-up approach to study finance, the research in the past half a century seemed to be centered around a top-down approach\(^1\), utilizing all summary statistics from the market, as led by the asset pricing literature which climaxed with the 2013 Nobel Memorial Prize in Economic Sciences. This thesis and the honorable title of “Dr.” shall prove as a certificate for my ambition and endeavor to push forward the research in finance via this bottom-up approach.

A perhaps further remote ideal prophesies, hopefully also in a self-fulfilling way, that the bottom-up and the top-down research directions will meet seamlessly someday. May this thesis be a humble, small step\(^2\) towards this ideal.

Bart Zhou Yueshen

Amsterdam, The Netherlands
May, 2014

\(^1\) The market microstructure literature has been growing, too. There are two surveys, Madhavan (2000) and Biais, Glosten, and Spatt (2005), to the best of my knowledge. Parlour and Seppi (2008) reviews the literature on limit order market, the dominant market structure in current times. The key expansion in the field is largely inspired by the rise of algorithmic and high-frequency trading in recent years. See Jones (2013) for a recent survey. O’Hara (2014) pushes forward the research agenda on high-frequency market microstructure.

\(^2\) “A humble, small step”: “跬步”——“不积跬步，无以至千里”；见《荀子·劝学》。
There are no competing financial interests that might be perceived to influence the analysis, the discussion, and/or the results of this thesis.
ROAD MAP

What this thesis looks at and where it heads
A proigious amount of financial securities’ trading volume is being generated across the world. These transactions are initiated by traders with heterogeneous motives, are conducted in distant markets of differing platforms, and are subject to a myriad of—sometimes contrasting—trading rules imposed by regulators. All these dimensions weave into a complex nature of financial securities’ trading and these multifarious trades constitute the very fundamental building blocks of the finance industry. It is, therefore, of paramount importance to understand the financial securities market—its functions, development, organization, design, and structure—and so is the aim of this thesis, as well as the future research that might bear its inspiration.

To march afar one always begins with a first step. Bearing the ambition to add to the ultimate understanding of financial securities’ trading, this thesis focuses on how some recent technological advancement have shaped the latest market structure. On the extensive margin, the technological advancement (together with the pro-competition attitude by the regulators) has had the terrain of financial markets trembled: The entry barrier of new market places has been significantly lowered, as witnessed by the sprouting of new high-tech trading venues that fragmented the market share of incumbent markets in the past decade. On the intensive margin, technology improvement has also triggered numerous transformations: Boosts in trading speed have enabled new trading strategies; new patterns from the market data are observed; adapted new players, computerized/algorithmic traders, join and start to dominate the arena; and new battlefields, alternative trading systems (electronic communication networks, crossing networks, dark pools, and the alike), are opened up and developed into different niches of trading needs. The subsequent three chapters of this thesis unveil the details of these recent technological changes, jointly contemplating on some new frictions therein and their implications, yet with respective own emphases on the various measures of market quality.

Chapter 2, based on Menkveld and Yueshen (2014a), is devoted to one of the most dramatic incidents in the history of financial securities trading: the Flash Crash. Both public and proprietary trade data on E-mini (S&P500 future) and SPY (S&P500 ETF) are used. The proprietary dataset allows the identification of the exact trades by a large fundamental seller, who allegedly ignited the nine percent evaporation of the entire U.S. stock market value (Dow Jones Industrial Average) in less than twenty minutes. It is shown that, however, this large seller chose to sell only mindfully and very limitedly during the period when the prices went through a free. Her price impact was magnified by other traders’ aggressive sells with a 300-millisecond delay. As the large seller kept selling (and turned even more aggressive) when the E-mini price bottomed, she reinforced the price pressure and paid excessive cost for transacting a large position. A calibration exercise suggests that her loss during the twenty-minute Flash Crash amounted to a quarter of her annual operating income.

The high-frequency empirical analysis points to the vulnerability of modern financial markets due to possible, unfortunate interaction among the market participants. Such a
crash inflicts a huge transaction cost on the fundamental participants of the market. It seems to be a legitimate, worrying concern that the possibility of incurring prohibitive transaction cost might shake investors’ confidence in the integrity of the current financial market.

Adapted from Menkveld and Yueshen (2014b), chapter 3 looks at the new middlemen who intermediate between end-users in financial securities trading. These middlemen are best thought of as represented by the high-frequency market makers that populate almost all electronic trading platforms nowadays. Compared to the existing literature on middlemen, the insight lies in the multiplicity of such middlemen: Do they interact with each other and, if so, how? What is the economic rationale? More importantly, what does such interaction imply for fundamental investors?

The chapter develops a tractable framework to address the above issues. Through the lens of the model, the effects of middlemen on welfare (allocative efficiency) of the market are scrutinized and identified exactly. In particular, both positive (liquidity provision) and negative (impairing information learning) channels are decomposed from the total marginal effect of an additional middlemen, and the sign of the net effect is pinned down to a parameter that measures how likely the middlemen are able to resell the asset.

It is shown that a fundamental investor might be “confused” by the market activity in that an observed transitory price pressure (due to middlemen multiplicity) could not be distinguished from a permanent price innovation (e.g. revelation of bad quality of the asset). When the former is the case, such inference problem will add to the overall transaction cost of the fundamental investor who inevitably trades on the price pressure and bears the associated risk inefficiently. The model implications appear reminiscent of the narrative of the Flash Crash (as outlined in chapter 2). It is argued that the mechanism discovered by the model could have played a significant role during the Flash Crash.

Chapter 4 is adapted from Yueshen (2014). It explores the optimal limit order submission strategies in an electronic limit order book, with the friction of random latencies between traders’ decision-making (submitting orders) and the decisions’ effectuation (orders being processed by the exchange server). The unknown latencies effectively queue the traders’ limit orders randomly (possibly dependent on traders’ different attributes, e.g., connection speed), and due to the time priority enforced by most of the real-world trading platforms, the associated profitability of the orders varies according to the queue realization.

Recent hardware upgrades, both on the exchange’s servers and on market participants’ end, have magnified the importance of such “queuing uncertainty”. The model developed in chapter 4 helps understand both the nature and the consequences of this friction. The novelty of the model, compared to the bulk of the literature on limit order market, lies in that instead of playing a (perfect information) sequential game, the agents move simultaneously, unable to observe the real-time market status. The model is applied to generate dynamics of liquidity provision, which complements the existing literature on the formation of a stable limit order book; to explain empirically observed phenomena (like so-called
CHAPTER 1. ROAD MAP

ghost/phantom liquidity in the market); and to weigh in on the debate of optimal market
design. It also provides a useful guidance on the empirical works that evaluate how trading
speed (high-frequency traders’ participation, server speed boosts, co-location service, etc.)
affects equilibrium order book depth.

Drawn from independent research papers, these three chapters combine to make a small
step toward a better understanding of financial securities’ trading. It is a mere attempt to
tackle a few of the many outstanding issues arising from the drastic development in the
field. There are yet many unexplored aspects, begging future research to clarify their roles
in resource allocation, asset valuation, information revelation, and so forth.

Such quests lie in the very heart of this thesis. It honors an idealized belief that these
small steps accumulated will push forward the boundary of the literature—the knowledge
base of human beings, enabling a broader horizon with a deeper understanding of the so-
ciety, the economy, and the finance world in particular. Thus is the direction to which this
thesis proudly points.
A large seller’s E-mini trades were in the center of the Flash Crash on May 6, 2010. This chapter contributes an empirical study of the event. A high frequency analysis is performed on the large seller’s trades and the trades of others in E-mini and SPY, two active securities to trade the S&P500. Price cointegration between them broke and prices collapsed about a minute before the E-mini halt, which marked a six percent drop from the pre-crash price level in E-mini. The large seller was relatively inactive in this period. Yet her trades, with a 300 millisecond delay, trigger aggressively selling by others. She sold most actively after the halt at highly-pressured prices. The price she paid for immediacy was disproportionately large, arguably due to the broken integration with other markets.
This chapter is based on Menkveld and Yueshen (2014a). Apart from the collaboration with Albert J. Menkveld, the discussions with and comments from Terrence Hendershott, Charles Jones, Emiliano Pagnotta, and Vincent van Kervel are gratefully acknowledged. The thanks also go to participants in an SEC conference call, conference/seminar participants at Euronext Paris/Toulouse University, Humboldt University, University College London, and the University of Gothenburg. This research is not possible without the generous data sponsorship by Eric Hunsader of Nanex and by Waddell & Reed. The VPIN estimates used in the this chapter is kindly shared by Maureen O’Hara.

2.1 Introduction

On May 6, 2010, U.S. equity indices declined by 5-6% and recovered, all in 30 minutes: the Flash Crash. The crash originated in the market for E-mini contracts, which are index futures on the S&P500. It spread rapidly to trading in other index products, but also to individual stocks (CFTC and SEC, 2010a,b). The crash echoed internationally as, for example, Canadian markets also crashed after approximately two minutes (IIROC, 2010).

There is a widespread concern that Flash Crash type events are the result of vulnerable electronic markets. Arguably the most damaging effect is that it might scare off market participants. For example, the Investment Company Institute (ICI) claimed that, as a result of the Flash Crash, “there have been five consecutive months of U.S. equity outflows.” The Joint CFTC-SEC Advisory Committee on Emerging Regulatory Issues\(^1\) writes: “While many factors led to the events of May 6, and different observers place different weights on the impact of each factor, the net effect of that day was a challenge to investors’ confidence in the markets” (Born et al., 2011, p.2). The SEC chairman voiced a similar concern in her opening statement at a special 2012 round-table triggered by Flash Crash: “…our concern is not whether a single firm might fail, but whether it causes collateral damage to investors and their confidence in the integrity and stability of our markets.”\(^2\)

Several empirical studies on the Flash Crash suggest the following narrative. A large seller initiated a sell program of 75,000 E-mini contracts worth approximately $4.1 billion (CFTC and SEC, 2010a). It did so in a market in which order flow grew more toxic by the hour (Easley, López de Prado, and O’Hara, 2012). Eventually, it triggered high-frequency traders’ selling and an E-mini price collapse (Kirilenko et al., 2011). It was followed by price declines in related markets, first index-tracking ETFs, then index securities themselves (Ben-David, Franzoni, and Moussawi, 2012). ETFs collapsed due to extreme liquidity deterioration, not a failure of market structure or the ETF product as such (Borkovec et al.,

\(^1\) The committee included academics as well as industry professionals. The academic members are Robert Engle, Maureen O’Hara, David Ruder, and Joseph Stiglitz.

2.1. INTRODUCTION

Hardest hit were ETFs that traded in fragmented markets, most affected by high-frequency trading and “inter-market sweep orders” (Madhavan, 2011).

This paper contributes to the literature in essentially two ways. First, it zooms in on the eye of the storm: trading by the large seller. An ultra-high frequency time series analysis relates the large seller’s net flow (i.e., market buy orders minus market sell orders) to other traders’ net flow, the E-mini price, and the net flow and price in SPY, an exchange-traded fund that is the most active alternative market to trade the S&P500. The results establish dynamic interrelationships i.e., do other variables respond to a shock in one of them? If so, with what time delay? And, what is the economic magnitude of the response? Such relationships will inform theoretical work on flash crashes as well as the regulatory debate.¹

The empirical analysis uses both public and proprietary data. The public dataset contains trades and price quotes for E-mini and SPY. The proprietary dataset consist of all E-mini trades done by the large seller in the Flash Crash period. These trades add up to the 75,000 contracts documented in CFTC and SEC (2010a). The granularity of the timestamps on all these data is 25 milliseconds.

A second contribution is that the paper develops an economic perspective. It calculates the price the large seller paid for immediacy. More importantly, it compares it to what a simple calibration exercise suggests is a “fair” price. The disproportionate price paid for immediacy is then related to the broken co-integration between the E-mini and the SPY price series. The remainder of the introduction reviews the main findings on both of these contributions.

Chronology of events. Trades of the large seller did not cause the steepest price declines in a direct manner. First, in the minute of the steepest price decline just before the halt, she contributed only 4% to the total net sells. Second, in this minute her trade intensity was low relative to what it was in earlier minutes. She decelerated. Third, only about half of the 75,000 contracts were sold before the halt. The remainder was sold mostly right after the halt at a time when the market started to recover.

The large seller trades however do seem related to the steepest price declines. In the minute before the halt, her occasional market sell order is followed, with a 300 millisecond delay, by a chain of aggressive sells by others. Initially, these sells amount to almost three times the size of the large seller order. Over time, the cumulative response grows to more than six times the initial order size. The price response to the large seller’s order is initially insignificant, but turns highly significant and large after 300 milliseconds. The long-run

¹The Flash Crash was by no means unique. Similar crashes hit for example the German DAX index (August 18, 2011 and April 17, 2013), the oil price (May 5, 2011), India’s National Stock Exchange index (October 5, 2012), the Anadarko stock (May 20, 2013), and the Procter and Gamble stock, (August 30, 2013). These recent crashes suggest a vulnerability in modern electronic markets, which worries regulators as well as the public at large.
cumulative response is 19 times higher than the benchmark i.e., the long-run response to
other traders’ net sells in that period. In fact, 43% of the price collapse can be explained by
net flow and almost half of it can be traced back to the large seller net flow. It is important
to note that the effect is indirect as it kicks in only after 300 milliseconds when other traders
aggressively sell and prices suddenly drop.

A maximum likelihood test reveals that price cointegration between E-mini and SPY
broke approximately one minute before the halt and was re-established about eight minutes
after the halt. The extraordinary E-mini conditions, most likely, caused the cointegration
to break. That is, before the break, each markets’ net flow has explanatory power for the
long-term (cointegrated) price. After the break, the SPY net flow no longer explains SPY
price. Instead, E-mini net flow explains both the E-mini price and the SPY price (which
became separate stochastic trends as cointegration broke).

There is a notable asymmetry in the way large seller trading affects market conditions
in the broken periods before and after the halt. After the halt, its net flow did not have a
disproportionate price impact. The E-mini net flow of other traders continued to impact both
the E-mini and the SPY prices. SPY net flow remained uninformative. One interpretation
is that the seller did not find outside customers before the halt. Instead it loaded up the
intermediation sector with its position and paid “price pressure” to compensate for inventory
cost. Kirilenko et al. (2011) provide supportive evidence as they report an order imbalance
between “fundamental buyers” and “fundamental sellers” before the halt, but not after. The
correspondence with the amount the large seller sold in this period is striking. The reported
imbalance before the halt is 49,665 - 83,599 = -33,934 contracts. The large seller sold
35,280 contracts in this period. After the halt, the imbalance is 110,369 - 110,177 = 192
contracts. In this period, the large seller sold 39,720 contracts. In sum, the halt became a
turning point as, after the halt, the large seller did find fundamental buyers in the E-mini
market. Before the halt, she did not.

Interpretively, one might say that the large seller was effectively “forced” to find fund-
damental buyers in the E-mini market and pay a price for it. She traded the remainder of
her order in the eight minutes of broken cointegration. Arbitrage forces seemed absent in
this period as the E-mini ask was about 100 basis points below the SPY bid. This price
wedge is an order of magnitude larger than the bid-ask spread in both markets at the time.
In effect, broken arbitrage barred her from finding counterparties (via arbitrageurs) in the
SPY market (and potentially other markets). It is tempting to interpret the price wedge as
the price the large seller paid for a broken market. At best, the wedge serves as an upper
bound as she could, most certainly, not have realized it in full given that her net flow in this
period is of non-trivial size relative to SPY net flow.\footnote{In this period, her net sells relative
to the standard deviation of SPY net flow is: $0.29 / $1.49 \approx 19.5\%$ (see table 2.2).}
2.1. INTRODUCTION

The large seller followed a 9% volume target and timed her trades to benefit from pockets of liquidity. A probit analysis relates her trade aggressiveness in the current 25 milliseconds to various lagged trade variables. The results are consistent with the 9% volume target as reported in CFTC and SEC (2010a). She turned more aggressive the more she fell short of that target based on her volume share in the preceding minute. The large seller’s strategy seemed to minimize the cost associated with taking liquidity. First, about half her trades were executed passively (it was her ask quote that was consumed by others as opposed to her consuming the bid quotes of others). Second, she only sold aggressively when, for example, bid depth was large or when the price had just moved up.

Assessment of the price paid for immediacy. Finally, a simple calibration exercise following Grossman and Miller (1988) suggests that the price the seller paid for immediacy was excessive. The model identifies under what conditions the price that was paid matches “immediacy production” by a zero-profit representative intermediary who bought the position from the large seller now in order to resell it to other buyers later. The price she paid could only be generated, for example, with very low capital in the intermediation sector (just enough to buy the $4.1 billion E-mini position), a very long holding period (three days), and a high relative risk aversion coefficient of 5.1

What trading friction then could have produced the high price the seller paid for immediacy? Two recent theoretical papers advance an explanation. Brunnermeier and Pedersen (2005) propose predatory trading by strategic speculators who sell along with a distressed seller in order to subsequently buy back the asset at a depressed price. The price overshoots and the liquidation value for the distressed seller is reduced. Chapter 3 advances an explanation based on imperfect learning. The large seller uses well-informed middlemen to reduce the adverse-selection problem she faces in a market with (potentially) informed buyers. Learning, however, is imperfect as inter-middlemen trades at a price discount and cannot be told apart from middleman-buyer trades at fundamental value. One result is that she sells “too much” at distressed prices (giving non-zero weight to the state that this is the fundamental price).

And, why did illiquidity spillover? One well-established channel is that illiquidity shocks propagate through an intermediation sector that supplies liquidity in multiple securities. A loss in one security (due to, for example, a mark-to-market loss on inventory) reduces wealth in the sector and, as a result, reduces its risk-bearing capacity.2 Gromb and Vayanos (2002) propose a model where arbitrageurs exploit price discrepancies of identical securities traded

---

1 Grossman and Miller (1988) is part of a classic microstructure literature often referred to as “inventory models”. In-depth surveys of microstructure theory can be found in O’Hara (1995), Madhavan (2000), Biais, Glosten, and Spatt (2005), and Foucault, Pagano, and Roell (2013).

in segmented markets and, by doing so, supply liquidity. As they need to collateralize in market separately, they might run out of “funding liquidity” (capital) when the price wedge widens. This effectively shuts them down and liquidity supply is reduced in both markets. In the same spirit, Brunnermeier and Pedersen (2009) model the feedback market liquidity can have on the margin requirement itself. They identify under what circumstances both become mutually reinforcing and “liquidity spirals” arise. Cespa and Foucault (2013) generate illiquidity contagion through a learning channel. One security suddenly turning illiquid reduces the information available to liquidity suppliers in correlated securities who, in turn, reduce their liquidity supply. This feeds back to liquidity suppliers in the first security. This feedback loop leads to a “liquidity crash.”

The remainder of the paper is organized as follows. Section 2.2 presents the data and performs basic data integrity checks. Section 2.3 revisits the chronology of events with a focus on large seller trading. Section 2.4 calculates the realized “price pressure” paid by the large seller and analyzes whether it was a “reasonable” price to pay for immediacy. Section 2.5 concludes.

2.2 Data

Data were supplied by Nanex, a firm that is specialized in low-latency (“real time”) distribution of trade-and-quote data to its clients. One dataset contains an ordered sequence of all May 6, 2010 order book events (trades and book changes) in the June-2010 E-mini contract. A second dataset contains similar information for an important index-tracker, SPY. All events carry a timestamp with a granularity of 25 milliseconds. They are recorded in Eastern Standard Time (EST). An important advantage of this data is that its creator is an information consolidator and distributor which guarantees consistency in event sequencing and timestamps. In other words, the data captures what low-latency participants saw and when they saw it.

The level of detail on market events differs across the E-mini and SPY dataset. The E-mini contract trades only at the Chicago Mercantile Exchange (CME). The data feed is very detailed as each order book change or trade becomes an event and is recorded in the database. SPY, however, trades on eight different exchanges: BATS, BOST, CBOE, CHIC, 1

1 As a matter of fact, the three-month LIBOR implied by the Eurodollar future jumped by about 15% in the seconds after the halt, from about 73 to 85 basis points. It recovered slowly in the minutes after. Borrowing suddenly became expensive, which is consistent with a funding liquidity friction.

2 SPY is the security symbol for “Spyder” (Standard & Poor’s Depository Receipts), an exchange-traded fund (listed on NYSE-ARCA) that tracks the S&P 500 index.

3 The Nanex system itself did not seem to suffer any delay in the Flash Crash period. One check that was done is to compare the timestamps on SPY quotes from NASDAQ ITCH to Nanex timestamps. The distance between the two timestamps was the same in the half hour Flash Crash period as compared to the earlier trading hours that day.
2.3. NARRATIVE OF THE FLASH CRASH

CINC, ISEX, NQES, and PACF. Its data feed is less detailed as only trades and changes in a market’s best bid or ask are recorded as events. For example, a depth change on the highest bid in BATS becomes an event, also when this bid is strictly lower than a bid from another exchange.

A separate, proprietary dataset was provided by Waddell & Reed (through Nanex). It contains all trades by the “large seller” who featured prominently in the CFTC-SEC report (CFTC and SEC, 2010a). All these trades can be matched with E-mini trade records in the public dataset. They are fully consistent with the CFTC-SEC description of the large seller’s trading: (i) they are all sells, (ii) they span a period of 20 minutes from 14:32 to 14:52, and (iii) they sum up to 75,000 contracts.

Some basic “data integrity” analysis was done and added as appendix 2.6.A. It documents that (i) there is no evidence on persistent, long-lived delays in quote reporting by exchanges, (ii) incidences of the ask price “touching” or “crossing” the bid are few, even in the half hour of the Flash Crash, and (iii) exchanges are allowed a 90 second delay on reporting trades which shows in the data; more than 90% of the trades, however, seem to be reported with, at most, a 100 millisecond delay.

2.3 Narrative of the Flash Crash

This section revisits the chronology of Flash Crash events with a focus on large seller trading. Section 2.3.1 takes a bird’s-eye view and compares her trading to that of others both before and after the E-mini halt. Section 2.3.2 performs an ultra-high frequency analysis on E-mini and SPY price cointegration and, based on the outcome, uses an appropriate econometric model to relate price changes to net flow. Finally, section 2.3.3 characterizes the trading strategy of the large seller by relating her trade aggressiveness to various lagged explanatory variables.

2.3.1 Exploratory analysis of the large seller’s trades

When did the large seller trade? Panel (a) of figure 2.1 illustrates E-mini trading in the Flash Crash period. It plots E-mini price and volume and, for each volume bar, indicates what part of it corresponds to large seller trading. It further reveals at what point the market hit a five-second trading halt, the “Stop Logic Functionality.” In this halt, “orders may be entered, modified or canceled but not concluded” (CME, 2010).

The graph leads to a couple of observations. First, it belies the straightforward explanation that the steep price drop was simply the result of an inactive, “illiquid” market. If anything, volume rose in the minute before the halt when the price collapsed. The price slid from 1150 to 1095 which is approximately a 5% drop. Second, this volume increase was not due to the large seller. As a matter of fact, she was less active in this minute compared
Figure 2.1: The Flash Crash on May 6, 2010. This figure plots volume and the midquote price series during the half hour period of the Flash Crash. The data is aggregated to one-second intervals. Panel (a) illustrates E-mini trading. Trade volume bars are stacked where the red/dark part is volume that involved the large seller and the green/light is volume that did not involve her. Panel (b) illustrates SPY trading. The timepoints for the E-mini SPY price cointegration breakdown and resumption are estimated in section 2.3.2.1.
to earlier minutes, both in absolute terms and relative to other traders’ volume. Third, she was most active in the three minutes after halt. In this period she sold most intensively and her volume was a larger part of total volume.

Panel (b) depicts SPY price and volume. The SPY market did not experience a trading halt but, for reference, the graph shows at what time the E-mini market hit a halt. Overall, the plot seems similar to the E-mini plot. The price collapse appears to be an almost perfect echo to what happened in the E-mini market, both in terms of size and speed. Also, volume rises in the minute leading up to the E-mini halt. The only striking difference is that SPY price recovery seems to be quicker. The price differential across markets and how it relates to trading, is analyzed in detail in section 2.3.2.

**With whom did the large seller trade?** Table 2.1 compares the large seller’s trading with the E-mini volume disaggregation results in Kirilenko et al. (2011). It leads to a couple of insights. First, in the 13 minutes of general price decline before the halt, there was a large imbalance in the order flow of fundamental buyers and sellers. Fundamental sellers sold 83,599 contracts, which is 33,934 more than what fundamental buyers bought in this period. This large imbalance is approximately the amount that the large seller sold in this period: 35,280. It seems that the large seller loaded up the intermediation sector with over 30,000 contracts. In other words, there did not seem to be enough fundamental buyers in the E-mini market for the amount that the large seller sold.

Second, in the 22 minutes of general price increase after the halt, there no longer is an imbalance between buyers and sellers in the E-mini market. Fundamental sellers sold 110,177 contracts, which is only 192 contracts less than what fundamental buyers. In this period, the large seller sold the remainder part of her order: 39,720 contracts. This time it seems that the large seller did find fundamental buyers in the E-mini market.

**2.3.2 High-frequency analysis of price changes and net order flows**

This section turns to a high-frequency analysis of price and order flow in the E-mini and SPY market. A standard tool in market microstructure to understand price change is to relate it to net flow. Such net flow is constructed by signing each trade according to whether it was buyer- or seller-initiated. A trade obtains a positive sign if it was a buyer who hit an ask quote and a negative sign if a seller hit a bid quote. The selection of the appropriate multivariate model, a vector autoregression (VAR) model or a vector error-correction model (VECM),

---

1 More precisely, none of the eight markets that trade SPY experienced a halt. For ease of exposition, the term “SPY market” is used throughout and refers to the aggregate across all of the eight markets.

2 The assumption is that in the classification scheme adopted by Kirilenko et al. (2011), the large seller would be labeled “fundamental seller.”
Table 2.1: Order flow by trader type in the Flash Crash period. This table presents average order flow (contracts bought minus contracts sold) by trader type in the Flash Crash period. It borrows heavily from (Kirilenko et al., 2011, table VII) who recognize two stages: a period of price decline until the E-mini halt (14:32:00-14:45:28) and a period of price recovery after the E-mini halt (14:45:33-15:08:00). Panel (a) documents order flow by trader type as reported by Kirilenko et al. (2011). In their classification scheme a market participant is a fundamental trader (buyer or seller) if her daily position change is greater than 15% of her daily volume. She is classified as an intermediary if the daily position change is smaller than 5% of her daily volume and if the account’s net holdings fluctuate within 1.5% of this volume. All intermediaries are ranked by the number of transactions and the top 7% is re-categorized as high-frequency traders (HFTs). Unlike other categories, intermediary/HFT categorization is based on the three days prior to the Flash Crash. A participant is a small trader if she traded less than 9 contracts that day. The remaining participants are labeled opportunistic traders. Panel (b) reports the order flow by the large seller. In total, she sold 75,000 contracts in this period.
depends on whether prices are cointegrated or not.\footnote{Chapters 9 and 10 of Hasbrouck (2007) provide an in-depth discussion of multivariate models in market microstructure.} In this particular case, it translates to whether or not the price differential across markets is stationary. This section first explores cointegration and then estimates the appropriate model.

### 2.3.2.1 Price cointegration

Figure 2.2 graphs the price differential across E-mini and SPY. It plots two price differentials: (i) the E-mini ask minus the SPY bid and (ii) the SPY bid minus the E-mini ask. To plot these two separate series is more informative than plotting the midquote differential (the midquote is defined as the average of the bid and ask quote). First, the midquote differential can be inferred from the two series as it is equal to their average. Second, the series themselves reveal whether or not arbitrage opportunities prevailed. This is the case when the first series turns negative or the second series turns positive (on the assumption that the trade fees are zero). Third, the two separate series are informative on the size of the bid-ask spread as their difference equals the sum of the E-mini and SPY bid-ask spread. This information reveals whether any arbitrage opportunity is large relative to the size of the spread.

The figure shows that the price differential seems to become non-stationary after the halt, accompanied by large arbitrage opportunities. The midquote differential gradually falls to about 3.5% in the early minutes after the halt. Only after approximately 10 minutes the price differential seems to become stationary again. The 3.5% is an order magnitude larger than the bid-ask spread in either market at the time. In fact, the full 3.5% seemed to be an arbitrage opportunity as the E-mini ask minus the SPY bid became -3.5%. Absent any further frictions, one could earn this price differential by aggressively buying E-mini and selling SPY.\footnote{Interactions with industry on this seemingly large arbitrage opportunity yielded several insights. First, given the enormous price decline and extremely low system response times, it was immediately clear that uncharted territory had been entered. Might orders execute on one leg of my arbitrage and not on the other? Perhaps they execute but get canceled \textit{ex-post} due to extreme market circumstances (as happened for equity trades)? Second, could price series be trusted? At some point delivery of the Dow Jones Index series was 80 seconds delayed (according to Nanex, see “The Annotated Flash Crash Diagram”, Financial Times, September 28, 2010). Third, margin requirements rose rapidly on steep increases in volatility. Netting across markets was impossible for most participants. Finally, the deal looked “too good to be true” (there must be something that others know and I do not know).}

The results of a breakpoint analysis reveal that cointegration broke about one minute before the halt and resumed eight minutes after the halt. Standard breakpoint estimation techniques yield estimates of the parameters of a \textit{stationary} model before and after a structural break, along with an estimate of when it occurred: the breakpoint. None of these techniques can be used off-the-shelf to estimate the potential cointegration breakpoints that
Figure 2.2: Price differential between E-mini and SPY. This figure plots two price differential series: (i) E-mini best ask minus SPY best bid and (ii) E-mini best bid minus SPY best ask. The three panels present these series for different time intervals. Panel (a) plots the entire trading day. Panel (b) plots the half hour Flash Crash period. Panel (c) zooms into a 20-second time window around the E-mini halt. The timepoints for the E-mini SPY price cointegration breakdown and resumption are estimated in section 2.3.2.1.
2.3. NARRATIVE OF THE FLASH CRASH

seem to characterize the price series in the Flash Crash period. There is, however, a relatively straightforward extension of these standard techniques that could be applied to the data sample. It is developed in appendix 2.6.B. This new approach finds that cointegration broke at 14:44:27.525 (1:00.625 before the halt) and recovered at 14:53:19.425 (7:46.325 after the halt) (see also panel (b) of figure 2.2 or figure 2.1). Note that cointegration broke down after the large seller started trading and just before the steep price decline. It recovered about one minute after the large seller’s last trade.

Based on these breakpoint estimates, together with the five-second trading halt, the Flash Crash is naturally divided into the following four subsamples:

- Period 1, 14:30:00.000 – 14:44:27.525, before the halt, E-mini and SPY cointegrated;
- Period 2, 14:44:27.525 – 14:45:28.150, before the halt, cointegration broken;
- Five-second trading halt;
- Period 3, 14:45:33.100 – 14:53:19.425, after the halt, cointegration broken;

All further analysis is done separately for each of these four periods as the nature of trade might differ substantially across these periods.

2.3.2.2 VAR/VECM analysis of price and net flow

The ultra-high frequency analysis of price and net flow dynamics in both markets recognizes the four distinct periods: pre-halt cointegrated, pre-halt broken, post-halt broken, post-halt cointegrated. A vector error correction model (VECM) is estimated for period 1 and 4 as the price is cointegrated across markets. In this case, both markets’ price series share a common random walk, often referred to as the “efficient price”. In period 2 and 3, price series correspond to two separate random walks and therefore requires estimation of a vector autoregressive (VAR) model. Both models are multivariate and thus describe price and flow interaction within a market, but also across markets.

Formally, the VECM/VAR model that is taken to the data is:

\[ y(t) = \sum_{i=0}^{k} \Phi_i y(t-i) + \left( p^E(t) - p^S(t) \right) \beta + \epsilon(t). \]

In the system, \( y(t) \) is a 5-by-1 vector that includes the following variables: large seller’s net flow, other traders’ net flow in E-Mini, net flow in SPY, E-Mini midquote return, and SPY midquote return. \( \Phi_i \) is a 5-by-5 coefficient matrix. In particular, \( \Phi_0 \) is set to such that only contemporaneous effects from order flows to returns are allowed, not the other way around.


(1) Hasbrouck, 2007). $\beta$ is a 5-by-1 coefficient vector that allows for an error-correction effect $(p^E(t) - p^S(t))$, where $p^E(t)$ and $p^S(t)$ are E-Mini and SPY log-prices. The $\beta$ is set to zero for periods 2 and 3 (cointegration broken i.e., the model becomes a VAR model) but estimated for periods 1 and 4 (the model is a true VECM model). Finally, $\epsilon(t)$ is a 5-by-1 vector of residuals.

The models are estimated based on 100 millisecond intervals in order to reduce the timestamp synchronicity issue that plagues trade reports (for a discussion see appendix 2.6.A). In each of the four periods, the number of lags was chosen according to the standard Akaike information criterion.

**Before the halt.** Before presenting the VAR/VECM estimation results, figure 2.3 zooms into the two minutes just before the halt. Panel (a) reveals that the large seller did trade in this period, although its last large burst of activity is centered around 14:44:15, only seconds ahead of the price collapse and the cointegration break. $^1$ She did continue to trade up until the halt, but at a much lower rate, both in absolute terms and relative to others. The graphs further reveals that, throughout, she executed both passively (her ask quote was consumed) and aggressively (she consumed other traders’ bid quotes). Panel (b) shows that, in this period, the large seller’s orders are large, but not extreme. They correspond to the largest orders traded by others, except for one large aggressive sell in the 14:44:15 burst of activity. It is an order of 737 contracts, which is about seven times the size of her largest order in this period. Yet, just ahead of it is an “extreme” 500-contract sell order by one of the other traders. The remainder of this section turns to econometric analysis in order to characterize the interaction of the large seller net flow, other traders’ net flow, and price.

Table 2.2 summarizes the VAR/VECM estimation results through impulse response functions. Panel (a) contains summary statistics on all trade and price variables. Panel (b) presents the E-mini and SPY price response to a $1 million impulse (about one standard deviation) in either large seller E-mini net flow ($q^E_{ls}$), other traders’ E-mini net flow ($q^E_{ot}$), or SPY net flow ($q^S_{ot}$). Panel (c) tests the symmetry of long-run order impacts on prices. Panel (c) shows how the net flow of each trader group responds to these impulses. The results lead to the following observations on trade dynamics.

$^1$ The burst consists of several aggressive sells, instead of a single large “bite” into the order book. None of these aggressive orders wiped out all the limit orders at the best price. For example, the largest order was for 737 contracts, executing against a resting depth of 1334 contracts at the best price at 14:44:14.850.
2.3. NARRATIVE OF THE FLASH CRASH

(a) Two minutes of trading before the halt

(b) Cross-sectional distribution of aggressive sell order size

Figure 2.3: Two minutes before the halt. Almost all of the Flash Crash price decline materialized in the minute before the halt. This figure illustrates it by zooming into the two minutes before the halt. The data is aggregated to one-second intervals. Panel (a) plots total volume. The stacked bars indicate large seller presence, both through her passive sells (her ask quote was consumed) and through her aggressive sells (she consumed a bid quote). This captures all her volume as she did not buy E-mini contracts during the Flash Crash. Panel (b) contains box plots to illustrate the cross-sectional distribution of the size of aggressive sells. It does so separately for large seller trades and other traders’ trades. A box is drawn from the 25% quantile to 75% quantile. A red bar depicts the median. Sells below the 25% and above the 75% quantiles are plotted individually and indicated by “+”.
## (a) Variable description

<table>
<thead>
<tr>
<th>notation</th>
<th>description</th>
<th>period 1</th>
<th>period 2</th>
<th>period 3</th>
<th>period 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>( q_{Els} )</td>
<td>large seller's E-mini net flow ($ mln)</td>
<td>-0.12</td>
<td>1.07</td>
<td>-0.04</td>
<td>0.44</td>
</tr>
<tr>
<td>( q_{Eot} )</td>
<td>other traders' E-Mini net flow ($mln)</td>
<td>-0.16</td>
<td>6.00</td>
<td>-0.87</td>
<td>7.21</td>
</tr>
<tr>
<td>( q^S )</td>
<td>SPY net flow in ($mln)</td>
<td>-0.01</td>
<td>0.97</td>
<td>0.09</td>
<td>1.41</td>
</tr>
<tr>
<td>( r^E )</td>
<td>E-Mini mid-quote return (bps)</td>
<td>-0.04</td>
<td>0.91</td>
<td>-0.54</td>
<td>5.09</td>
</tr>
<tr>
<td>( r^S )</td>
<td>SPY mid-quote return (bps)</td>
<td>-0.04</td>
<td>0.70</td>
<td>-0.45</td>
<td>5.25</td>
</tr>
</tbody>
</table>

### Table 2.2: VAR/VECM analysis of return and net order flow in E-mini and SPY.

Panel (a) summarizes the variables used in the VAR/VECM analysis. The variables are calculated based on 100-millisecond intervals. The net order flow is converted to a dollar value based on the security’s midquote value at 9:30 on May 6. VECM is estimated for period 1 and 4 (E-mini and SPY prices are cointegrated). VAR model is estimated for period 2 and 3 (broken cointegration). Panel (b) and (d) summarize the estimation results by presenting impulse response functions for price and other traders’ order flow, respectively. Panel (c) presents the result of a test on whether different net flow impulses have the same long-run price impact. The impulse is a net sell of $1 million (which is about the standard deviation of the large seller’s net flow). The impulse responses allow for a contemporaneous effect only of an order flow impulse on a security’s return, not the other way around. Statistical significance is established through simulations.
2.3. NARRATIVE OF THE FLASH CRASH

(b) Price impulse response functions

<table>
<thead>
<tr>
<th>#obs.</th>
<th>shock in</th>
<th>shock size ($1 mln)</th>
<th>response of</th>
<th>cumulative price changes (bps)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>E-Mini</td>
<td>100ms</td>
</tr>
<tr>
<td>period 1 (before halt, cointegrated)</td>
<td>$q_0^E(t)$</td>
<td>-1.0</td>
<td>E-Mini</td>
<td>-0.03*** -0.03** 0.03*** -0.03 -0.03 -0.03 -0.03 -0.03 -0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SPY</td>
<td>-0.01** -0.02*** -0.02*** -0.03*** -0.03*** -0.03 -0.03 -0.03 -0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$q_0^E(t)$</td>
<td>8.674</td>
<td>E-Mini</td>
<td>-0.03*** -0.03** -0.03*** -0.03*** -0.03*** -0.03*** -0.03*** -0.03*** -0.03***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SPY</td>
<td>-0.01*** -0.02*** -0.03*** -0.03*** -0.03*** -0.03*** -0.03*** -0.03*** -0.03***</td>
</tr>
<tr>
<td></td>
<td>$q^E(t)$</td>
<td>-1.0</td>
<td>E-Mini</td>
<td>-0.06*** -0.06*** -0.06*** -0.07*** -0.07*** -0.10*** -0.10*** -0.10*** -0.10***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SPY</td>
<td>-0.07*** -0.08*** -0.09*** -0.09*** -0.09*** -0.09*** -0.10*** -0.10*** -0.10***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>period 2 (before halt, cointegration broken)</td>
<td>$q_0^E(t)$</td>
<td>-1.0</td>
<td>E-Mini</td>
<td>-0.50 -0.27 -2.36*** -2.54*** -2.84*** -2.42*** -2.72*** -2.72***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SPY</td>
<td>0.03 -0.42 -0.23 -1.64*** -1.84*** -2.12*** -1.82***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$q_0^E(t)$</td>
<td>605</td>
<td>E-Mini</td>
<td>-0.06* -0.09 -0.10 -0.08 -0.05 -0.14* -0.14* -0.14*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SPY</td>
<td>-0.05*** -0.07*** -0.08*** -0.09*** -0.08 -0.11* -0.13* -0.13*</td>
</tr>
<tr>
<td></td>
<td>$q^E(t)$</td>
<td>-1.0</td>
<td>E-Mini</td>
<td>0.12 0.03 0.03 -0.01 -0.03 0.41 0.30 0.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SPY</td>
<td>0.03 -0.00 0.09 -0.07 -0.03 0.30 0.23 0.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>period 3 (after halt, cointegration broken)</td>
<td>$q_0^E(t)$</td>
<td>-1.0</td>
<td>E-Mini</td>
<td>0.05 0.05 -0.10* 0.14** 0.07 0.00 -0.10 -0.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SPY</td>
<td>-0.02*** -0.01** -0.02*** -0.03*** -0.00 -0.01 -0.06 -0.06</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$q_0^E(t)$</td>
<td>4.662</td>
<td>E-Mini</td>
<td>-0.08*** -0.10*** -0.10*** -0.12*** -0.13*** -0.13*** -0.11***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SPY</td>
<td>-0.02*** -0.03*** -0.04*** -0.04*** -0.05*** -0.06*** -0.06***</td>
</tr>
<tr>
<td></td>
<td>$q^E(t)$</td>
<td>-1.0</td>
<td>E-Mini</td>
<td>0.01 0.01 0.01 -0.02 -0.04 0.00 0.02 0.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SPY</td>
<td>-0.01 -0.00 0.03 0.12 0.06 -0.08 -0.03 -0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>period 4 (after halt, cointegrated)</td>
<td>$q_0^E(t)$</td>
<td>-1.0</td>
<td>E-Mini</td>
<td>-0.03*** -0.03*** -0.03*** -0.03*** -0.03*** -0.03*** -0.03***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SPY</td>
<td>-0.03*** -0.04*** -0.04*** -0.04*** -0.04*** -0.04*** -0.04***</td>
</tr>
<tr>
<td></td>
<td>$q^E(t)$</td>
<td>-1.0</td>
<td>E-Mini</td>
<td>-0.00 0.00 0.00 -0.00 0.00 0.00 0.00 0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SPY</td>
<td>-0.00 -0.00 0.00 -0.00 -0.00 -0.00 -0.00 -0.00</td>
</tr>
</tbody>
</table>

* ** *** Significant, respectively, at 10%, 5%, and 1%. All tests are two sided.

(c) Tests on the long-run price impact

<table>
<thead>
<tr>
<th>absolute long-run impact difference</th>
<th>period 1 (cointegrated)</th>
<th>period 2 (broken)</th>
<th>period 3 (broken)</th>
<th>period 4 (no large seller)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td>q_0^E(t) - \bar{q}_0^E(t) - \ldots - q_0^E(t)</td>
<td>$</td>
<td>-0.01 (-0.41)</td>
<td>2.58* (2.48)</td>
</tr>
<tr>
<td>$</td>
<td>\bar{q}_0^E(t) - \bar{q}_0^E(t) - \ldots - q_0^E(t)</td>
<td>$</td>
<td>-0.01 (-0.41)</td>
<td>1.69* (2.32)</td>
</tr>
</tbody>
</table>

* ** *** Significant, respectively, at 10%, 5%, and 1%. All tests are two sided. T-statistics in brackets.
Table 2.2 continued…

(d) Net flow impulse response functions

<table>
<thead>
<tr>
<th>#obs.</th>
<th>shock in</th>
<th>shock size</th>
<th>response of</th>
<th>cumulative order flow response (1 mn)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>period 1 (before halt, cointegrated)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$q_{61}(t)$</td>
<td>-1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$q_E$</td>
<td>-1.11*** -1.23*** -1.25*** -1.27*** -1.30*** -1.38*** -1.41*** -1.44***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$q_{61}^E$</td>
<td>-0.07 0.07 0.16 0.17 0.12 -0.06 -0.09 -0.09</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$q_S$</td>
<td>-0.02 -0.02 0.00 -0.00 -0.02 -0.01 -0.01 -0.00</td>
</tr>
<tr>
<td>8,674</td>
<td></td>
<td></td>
<td>$q_{61}^S$</td>
<td>0.00 0.00 0.00 0.00 0.00 0.01 0.01 0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$q$</td>
<td>-1.10*** -1.10*** -1.11*** -1.16*** -1.19*** -1.29*** -1.33*** -1.33***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.01** -0.01* -0.01** -0.01*** -0.01*** -0.01 -0.01 -0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.00 0.00 0.00 0.00 0.00 0.01 0.02 0.03 0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.12 0.04 0.04 0.03 0.00 -0.30 -0.39 -0.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-1.04*** -1.05*** -1.09*** -1.12*** -1.13*** -1.16*** -1.16*** -1.16***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$q^E(t)$</td>
<td>-1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$q_{61}^E$</td>
<td>-0.99*** -1.05*** -1.11*** -1.13*** -1.15*** -1.27*** -1.36*** -1.36***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$q_{61}^E$</td>
<td>-0.39 -0.20 -2.91*** -4.18*** -4.58*** -5.41*** -6.47*** -6.47***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$q_S$</td>
<td>0.15 0.30 0.36 0.40 0.47 0.27 0.19 0.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$q_{61}^S$</td>
<td>0.00*** 0.01*** 0.01* 0.01 0.01 0.00 0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$q$</td>
<td>-1.00*** -0.95*** -0.94*** -0.97*** -0.94*** -1.04*** -1.09*** -1.09***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.03*** 0.04*** 0.05* 0.05** 0.06** 0.06 0.06 0.06</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.10 -0.08 -0.06 -0.10 -0.13 0.56 0.61 0.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-1.17*** -1.17*** -1.19*** -1.20*** -1.16*** -1.15*** -1.15*** -1.15***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$q^S(t)$</td>
<td>-1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$q_{61}^S$</td>
<td>-1.00*** -1.12*** -1.20*** -1.28*** -1.43*** -1.77*** -2.67*** -2.72***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$q_S$</td>
<td>-0.01 0.00 -0.05 -0.04 -0.05 0.13 0.42 0.44</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$q_{61}^S$</td>
<td>-0.00 0.00 0.02 0.01 0.02 -0.03 -0.17* -0.17*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$q$</td>
<td>0.00 0.00 0.00 0.00 0.01 0.01 0.02 0.03 0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-1.08*** -1.08*** -1.08*** -1.11*** -1.08*** -1.11*** -1.09*** -1.09***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.00 0.01* 0.01 0.01 0.01 0.01 0.01 0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.00 -0.01 -0.01 -0.01 -0.03 -0.03 -0.10 -0.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.05 0.06 0.03 -0.09 -0.11 0.03 0.12 0.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-1.05*** -1.10*** -1.10*** -1.15*** -1.18*** -1.22*** -1.27*** -1.27***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$q^S(t)$</td>
<td>-1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$q_{61}^S$</td>
<td>-1.05*** -1.07*** -1.09*** -1.09*** -1.09*** -1.10*** -1.11*** -1.11***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$q_S$</td>
<td>0.00 -0.00 -0.00 -0.00 -0.00 -0.00 0.00 0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$q_{61}^S$</td>
<td>0.00 0.03 0.01 0.01 0.01 0.01 0.01 0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$q$</td>
<td>-1.00*** -1.01*** -1.02*** -1.02*** -1.02*** -1.02*** -1.02*** -1.02***</td>
</tr>
</tbody>
</table>

* ** *** Significant, respectively, at 10%, 5%, and 1%. All tests are two sided.
Before the halt, the average E-mini net flow was negative for both the large seller and the other traders. This is not surprising for the large seller since (i) she did not buy E-mini in the Flash Crash period (only sold it) and (ii) figure 2.3 revealed that some of her trades were aggressive. The average negative net flow of others is more surprising. Apparently, the total amount that they sold aggressively was larger than what they bought aggressively. Before the cointegration break, the large seller’s aggressive sell intensity is about the same size as that of the other traders: $120,000 versus $160,000 per 100 milliseconds, respectively. After the break, the large seller reduced her intensity by 66% to $40,000 per 100 milliseconds whereas the other traders increased their intensity by 444% to $870,000 per 100 milliseconds. Her net sell rate was therefore only $40,000 / $910,000 ≈ 4% of the overall net sell rate. Coincidentally, the price run down in this period is about 10 times larger than what it was before break: 0.54 and 0.04 basis points per 100 milliseconds, respectively.

The large seller’s trades have the highest long-run price impact in period 2. The impact is an order of magnitude larger than her impact in other periods or other traders’ impact in all periods for that matter. A $1 million aggressive sell impulse (roughly one standard deviation) by the large seller has a long-term price impact on E-mini of 2.72 basis points, most if it kicking in after 300 milliseconds. This impact is extraordinary as similar flow by the large seller in the other three periods yields a long-term impact of maximum 0.11 basis points and other traders’ flow has a maximum impact of 0.14 basis points (obtained in period 2). Statistical tests reveal that indeed the large seller’s long-run impact is significantly higher in period 2, but not in the other periods.\(^1\)

The large seller’s net flow contributes almost half of the price drop that is predicted by net flow in period 2. The rate at which the price drops in this period is 0.54 basis points per 100 milliseconds. The average net sell rate of other traders’ implies a rate of 0.87 × 0.14 ≈ 0.12 basis points per 100 milliseconds. The large seller’s implied rate is 0.04 × 2.72 ≈ 0.11 basis points. Added up, the implied rate “explains” \((0.11 + 0.12) / 0.54 \approx 43\%\) of the observed rate at which the price drops in this period.

Further analysis of this disproportionate response to large seller flow reveals that most of it appears with a 300 millisecond delay at a time there is also a strong aggressive sell response by other traders. The immediate response of other traders (i.e., with a 100 milliseconds delay) to a large seller flow impulse is insignificant and “only” 0.39 the size of the impulse. It however suddenly swells to a significant (cumulative) response of 2.91 after 300 milliseconds. A similar pattern is observed for the price response. It is insignificant and small initially but, after 300 milliseconds, it turns significant and large: an initial response

\(^1\)This disproportionate response is not driven by an asymmetry in the impact of other traders’ net buy flow or net sell flow. It remains disproportionately large relative to the impact of \(q_{ot}^+\) in an extended model with \(q_{ot}^+ = \max(0, q_{ot}^-)\) and \(q_{ot}^- = \min(0, q_{ot}^+)\) as separate explanatory variables. This model is not presented in the main text as the null hypothesis that the long-run price impact of these two variables is the same cannot be rejected.
(after 100 milliseconds) of -0.50 basis points becomes -2.36 basis points. The response is “permanent” as after 10 seconds it seems to settle at -2.72 basis points.

Figure 2.4 illustrates the chain of events suggested by the period 2 impulse response functions. It depicts two instances in this period where the large seller aggressively sells without an initial price response. Then, after several 100 millisecond intervals, other traders aggressively sell and the price declines steeply.

In terms of market interaction, the SPY price seems to respond to E-mini flow with delay whereas there is no such delay in E-mini price response to SPY flow. In period 1 for example the SPY cumulative price response to an E-mini other traders’ net sell impulse is -0.01 basis points in the first 100 milliseconds and doubles the next 100 milliseconds to -0.02. The E-mini price response to a SPY net sell impulse is -0.06 after 100 milliseconds and remains at about that level after 200 milliseconds: -0.06 basis points. This finding is perhaps not that surprising given that E-mini is the more active market. On May 6 its volume is about four times larger than the aggregate SPY volume. This finding is consistent with Hasbrouck (2003) who focuses on price discovery in E-mini, SPY, and a regular S&P 500 futures contract traded in the pit. For a March 2000 sample he documents that the E-mini information share ranges from 90.9% to 93.3%.

Notably, this E-mini lead in price discovery pertains also to period 2 when cointegration is broken and the SPY price is not disciplined to stay within a “stationary” distance from the E-mini price. In fact, SPY net flow is no longer significant as an explanatory variable for SPY price changes. The latter is only significantly explained by E-mini net flow, both its large seller component and the other traders’ component. The SPY price response function closely resembles the E-mini price response function with two main differences. First, its response to a large seller impulse materializes only after 400 milliseconds whereas the E-mini price response is with a 300 millisecond delay. Second, the response function is scaled down by about two-thirds which could explain the cointegration break.

After the halt. In period 3, the broken cointegration period immediately after the halt, the large seller sold aggressively and other traders’ net flow was positive on average. In this period, the large seller’s net sell intensity was $290,000 per 100 milliseconds, which is more than double her pre-halt intensity. This time, however, other traders’ net flow was positive and averages $360,000 per 100 milliseconds. The aggregate net flow therefore became positive in this period: $70,000 per 100 milliseconds. The E-mini and SPY price both increased in this period at an average rate of 0.11 and 0.10 basis points per 100 milliseconds, respectively. Compared to period 2, E-mini net flow continues to impact both price series and SPY net flow remains insignificant. Large seller flow, though, turns insignificant. In summary, it seems that the halt did break the downward spiral of a large seller net sell impulse followed by a larger net sell response by other traders. Instead, other traders seem to recognize that prices are pressured and aggressively buy to earn the “liquidity premium”.
2.3. NARRATIVE OF THE FLASH CRASH

Figure 2.4: The large seller’s order impact in period 2 (one minute before the halt). This figure illustrates the impact of the large seller’s net order flow in period 2, the minute of broken cointegration before the halt. The two panels depict trading at a high frequency: 100 millisecond intervals. The graphs are meant to illustrate the results of the VAR model that was estimated for this period (see table 2.2).
Period 4 starts about a minute after the large seller finished selling and trade looks much like what it was in period 1 except that SPY net flow is insignificant as an explanatory variable for long-run price changes. Cointegration recovered and the long-run E-mini and SPY price response to net flow impulses became equal again. The response to an E-mini net sell impulse is -0.03 basis points after 10 seconds which is 26% less than what it was in period 1. Figure 2.1 illustrates that in this period also the general price level had recovered to what it was for the most part of period 1. It seems that in period 4 the markets had fully recovered.

2.3.3 Characterization of the large seller’s algorithm

This subsection characterizes the large seller’s trading, first through summary statistics and then through a conditional analysis of its aggressive sell size decision.

2.3.3.1 The large seller’s size in the market

**Large seller’s trade size.** The trade size distributions plotted in figure 2.5 reveal that the large seller trades in sizes that are not larger than the trades done by others. She does however trade large sizes relatively more often. The linear scale graphs show that about one fifth of her trades are larger than 50 contracts whereas for other traders this frequency never exceeds three percent. The logarithmic graphs reveal that large trades in this 50+ bucket are still smaller than the largest trades observed for others.

**Large seller’s volume share.** The large seller volume share plotted in figure 2.6 reveals that she indeed followed the reported 9% execution target. (CFTC and SEC, 2010a, p.2) states:

“This large fundamental trader chose to execute this sell program via an automated execution algorithm (‘Sell Algorithm’) that was programmed to feed orders into the June 2010 E-mini market to target an execution rate set to 9% of the trading volume calculated over the previous minute, but without regard to price or time.” [emphasis added]

The average execution rate seems to be 9% although, in particular in period 1 and 2, she often either fell short or overshot it by a couple of percentage points. The only notable exception is her extremely high trade rate in the minutes after the halt when her volume share touches 25%. The overall trade rate that was reported in figure 2.1 shows that this volume share increase was driven by the large seller trade rate increase as opposed to a drop in other traders’ trade rate. Figure 2.6 further shows that at each point in time the large seller executes passively and aggressively at about the same rate. Also here, however, the minutes
2.3. NARRATIVE OF THE FLASH CRASH

(a) Period 1 (14:30:00-14:44:27.525), before the halt and cointegrated

(b) Period 2 (14:44:27.525-14:45:28.150), before the halt and cointegration broken

(c) Period 3 (14:45:33.100-14:53:19.425), after the halt and cointegration broken

Figure 2.5: Trade size distribution. This figure depicts the distribution of the size of aggressive sell trades by the large seller and by the other traders. The three panels correspond to the three periods in which the large seller was present. In each panel, the left graph plots a histogram with a 50+ bucket for very large trades. The right graph provides further insight into these large trades by using a log scale on both axes.
after the halt are exceptional as her volume share spike is solely the result of an increase in the rate of aggressive orders. She seems to value immediacy particularly highly in these post-halt minutes.

2.3.3.2 Predictors of large seller aggressiveness

The CFTC/SEC quote referenced above (page 26) makes it sound as if volume share was the seller’s only target. It explicitly states that she disregarded price and time. The remainder of this section relates her action choice to her volume share in the preceding minute, but also to other trade variables based on price and time (where “time” is interpreted broadly as market conditions at the time of the seller action). The natural choice for dependent variable is the size of her aggressive sell order as she has full control over it (as opposed to passive orders that require aggressive buyers for execution; note that data is available only on her trades, not her orders).

**Ordered probit.** An appropriate model for the conditional analysis of the large seller’s aggressiveness is ordered probit. Aggressiveness is measured by the size of the large seller’s aggressive trades. Each observed trade size is assigned to one of four bins that are constructed based on the empirical distribution’s quartiles. By construction, all bins contain

![Figure 2.6: Large seller’s volume share.](image-url) Inspired by the CFTC-SEC report (CFTC and SEC, 2010a), this figure plots the large seller’s volume share for each minute in the Flash Crash half hour. The minutes considered are all those that exist on a 25-millisecond time grid as this is the highest frequency available in the data. The report claims that the large seller followed a 9% target.
(almost) the same number of observations. Ordered probit then “regresses” an observation’s bin number (where four is the bin with largest order sizes) on various explanatory variables.

Ordered probit is preferred over linear regression as it (i) allows for series that are non-negative, (ii) it allows for non-linearity (by varying the distance between thresholds), and (iii) it is more robust to outliers. Consider the last point. Linear regression will aim to fit a single, extremely large draw on the left-hand side variable as not doing so is extremely costly in terms of the “least squares” criterion. Ordered probit, on the other hand, first assigns observations to categories and then tries to fit the category “number” to a set of explanatory variables. In spite of all objections, the results of a linear regression are similar to those of the probit analysis.

**Explanatory variables.** The level of order aggressiveness that the large seller chose is related to a set of lagged explanatory variables. The set includes the standard variables featured in earlier work on order aggressiveness (e.g., Griffiths et al., 2000 and Ranaldo, 2004): price return, relative bid-ask spread, and depths at the best bid and best ask quote. Also included is realized volatility and the variance ratio as they are standard measures in the microstructure literature to characterize trading conditions. The variance ratio measures the relative size of price reversals as it relates high-frequency return variance to low-frequency return variance. The more the ratio exceeds one, the more of the price variation is due to reversals. Finally, a volume target error variable is added to capture a potential feedback loop in her trading in order to reach an overall 9% volume share target. It is defined as the seller’s volume in the preceding minute minus what it should be, given a 9% volume target. This explanatory variable is preferred over realized volume share minus 9% because the dependent variable is a size variable (not a percentage variable).

A grid search over all possible historical time windows reveals that the seller’s trade aggressiveness is most responsive to explanatory variables based on a three-second most recent history. The time window, however, is only relevant for those variables that require an interval to compute (e.g., volatility) in contrast to end-point variables (e.g., bid-ask spread). Figure 2.7 plots the probit log-likelihood for all windows where the window size varies from 1 second to 5 seconds and the end point is -150 milliseconds to no millisecond delay. These plots can only be made for period 1 and 3, as period 2 has too few observations to estimate the model (only 37 aggressive sells by the large seller). The plots for both periods show that the large seller’s algorithm is “low latency” as for any starting point, the highest likelihood is obtained at the no delay end point. The highest likelihood is obtained for the starting point of minus three seconds.

**Estimation results.** Table 2.3 presents the results of ordered probit model estimation. These results lead to the following observations. First, in period 1 the large seller is signifi-
(a) Variable description and summary statistics

<table>
<thead>
<tr>
<th>variable</th>
<th>description</th>
<th>period 1</th>
<th>period 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>vlm target error</td>
<td>large seller's volume in previous minute minus 9% of total volume (1,000 contracts)</td>
<td>-0.23</td>
<td>0.64</td>
</tr>
<tr>
<td>* midquote return</td>
<td>E-Mini mid-quote return (bps)</td>
<td>1.85</td>
<td>8.56</td>
</tr>
<tr>
<td>* ls passive vlm</td>
<td>the large seller's passive sell volume (1,000 contracts)</td>
<td>0.15</td>
<td>0.14</td>
</tr>
<tr>
<td>* sell vlm excl. ls</td>
<td>total aggressive sell volume, excl. the large seller (1,000 contracts)</td>
<td>0.75</td>
<td>1.02</td>
</tr>
<tr>
<td>* buy vlm. excl. ls</td>
<td>total aggressive buy volume, excl. those bought from the large seller (1,000 contracts)</td>
<td>0.95</td>
<td>1.18</td>
</tr>
<tr>
<td>relative spread</td>
<td>bid-ask spread divided by midquote (bps)</td>
<td>2.47</td>
<td>8.20</td>
</tr>
<tr>
<td>ask depth</td>
<td>depth at the best ask price (1 contract)</td>
<td>288.4</td>
<td>58.2</td>
</tr>
<tr>
<td>bid depth</td>
<td>depth at the best bid price (1 contract)</td>
<td>170.5</td>
<td>64.6</td>
</tr>
<tr>
<td>* volatility</td>
<td>realized volatility of based on midquote return (bps)</td>
<td>0.54</td>
<td>3.77</td>
</tr>
<tr>
<td>* variance ratio</td>
<td>$2 \times \text{var}<em>{hf}/(\text{var}</em>{hf} + \text{var}<em>{lf})$, where $\text{var}</em>{hf}$ is high-frequency (25ms) and $\text{var}_{lf}$ is low-frequency (250ms) midquote return variance</td>
<td>1.21</td>
<td>1.31</td>
</tr>
</tbody>
</table>

* Flow variable calculated based on a three second window just before the trade.

(b) Ordered probit estimation result

<table>
<thead>
<tr>
<th>variable</th>
<th>Average effect</th>
<th>period 1</th>
<th>period 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>vlm target error</td>
<td>$z$-stat.</td>
<td>$\Delta P(y=1)$</td>
<td>$\Delta P(y=2)$</td>
</tr>
<tr>
<td>-3.61***</td>
<td>-0.183</td>
<td>0.010</td>
<td>-0.075</td>
</tr>
<tr>
<td>midquote return</td>
<td>3.31***</td>
<td>-0.179</td>
<td>-0.040</td>
</tr>
<tr>
<td>ls passive vlm</td>
<td>0.62</td>
<td>-0.030</td>
<td>-0.005</td>
</tr>
<tr>
<td>sell vlm excl. ls</td>
<td>0.06</td>
<td>-0.004</td>
<td>-0.001</td>
</tr>
<tr>
<td>buy vlm. excl. ls</td>
<td>-1.92*</td>
<td>0.137</td>
<td>0.011</td>
</tr>
<tr>
<td>relative spread</td>
<td>0.18</td>
<td>-0.008</td>
<td>-0.001</td>
</tr>
<tr>
<td>ask depth</td>
<td>2.85***</td>
<td>-0.118</td>
<td>-0.023</td>
</tr>
<tr>
<td>bid depth</td>
<td>4.45***</td>
<td>-0.185</td>
<td>-0.042</td>
</tr>
<tr>
<td>volatility</td>
<td>-0.62</td>
<td>0.045</td>
<td>0.005</td>
</tr>
<tr>
<td>variance ratio</td>
<td>-1.89*</td>
<td>0.082</td>
<td>0.008</td>
</tr>
</tbody>
</table>

**lnL** | $P(y=1)$ | $P(y=2)$ | $P(y=3)$ | $P(y=4)$ |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>-960.6</td>
<td>0.36</td>
<td>0.17</td>
<td>0.22</td>
<td>0.25</td>
</tr>
</tbody>
</table>

-9103.0         | 0.26     | 0.24     | 0.25     | 0.25     |

*, **, *** Significant, respectively, at 10%, 5%, and 1%. All tests are two sided.

Table 2.3: Ordered probit analysis of the large seller’s trade aggressiveness. This table “regresses” the size of aggressive sell order by the large seller on various explanatory variables. These variables are calculated based on a three second window just ahead of the trade (this window is discovered through maximum likelihood, see figure 2.7). Flow variables (e.g., volatility) are calculated from the entire window. Stock variables (e.g., bid depth) are calculated from the window’s endpoint. Panel (a) describes the variables. Panel (b) presents the results of the ordered probit regression for periods 1 and 3 (too few observations in period 2). The probit model assigns the large seller’s trade size to four bins that are constructed based on the quartiles (all bins are therefore of approximately equal size). Bin 1 is the smallest and bin 4 is the largest.
cantly less aggressive when the volume target error is large. The effect is large economically as a one standard deviation change in the error reduces the likelihood of a Q4 (4th-quartile) size trade (large sell) by about 20.5%. In period 2 this coefficient is not significant. This is consistent with figure 2.6 that shows that in the minutes after the halt she aggressively traded far beyond the 9% threshold.

Second, the large seller does take price change into account: She sells more aggressively after an upturn. The results are both statistically and economically significant. For example, a one standard deviation higher midquote return makes a Q4 sell 24.0% more likely in period 1 and 14.8% more likely in period 3.

Third, in terms of liquidity supply in the book, the large seller becomes more aggressive if the bid depth is larger, the ask depth is larger, and the spread is larger. The strongest finding is increased aggressiveness on large bid depth as it is statistically significant in both periods. Its economic size is substantial as a one standard deviation change makes a Q4 size trade 25.0% more likely in period 1. In period 3, this likelihood is increased by 29.8%. The result is intuitive. The size of her aggressive sells is larger if more is supplied on the other side of the market. The other two results are weaker. Ask depth is significant only in period 1 and the spread only in period 3. Also economically the effect is weaker as Q4 size trade rises by “only” 15.4% and 17.1%, respectively. These results are less intuitive. Perhaps a lot of ask side depth implies that her limit sell orders in the book face more competition and she therefore needs to rely more on aggressive orders to maintain a target selling rate. The spread result is counter-intuitive. Further study revealed that it is driven entirely by the early part of period 3, the period in which she sold aggressively and her volume share approached 25% (until about 14:46:48, see figure 2.6).

Finally, the large seller becomes less aggressive when the market turns more volatile or when more of the volatility is due to transitory price effects (“pricing errors”). The general volatility effect is only significant in period 3. Its effect is large as a one standard deviation change makes a Q4 size trade 22.3% more likely. The transitory volatility effect is significant in both periods but its economic size is relatively modest. A one standard deviation increase in the variance ratio increases Q4 size sells by 9.6% and 11.8% in period 1 and 3, respectively.

Illustration of the results. Figure 2.8 illustrates that the large seller actively times her trades as suggested by the probit findings. She aggressively sells after the midquote increases for a couple of seconds just before 14:35:30, and she refrains from selling when the price slides in the earlier part of that minute.

Figure 2.9 suggests that the large seller’s relative presence in the market co-moves negatively with flow toxicity. This finding is consistent with strategic trading: she sells passively during upturns (her limit sell orders are taken out), sells aggressively right after an upturn, and does not trade in downturns (see Figure 2.8). She therefore is absent in the “informa-
CHAPTER 2. THE FLASH CRASH

Figure 2.7: Trade history window that the large seller is most sensitive to. This figure illustrates to what time window of trade history the large seller’s strategy is most sensitive to. It graphs contour plots of log-likelihoods associated with probit “regressions” of her aggressive order size choice on various explanatory variables. The plots show how the log-likelihood changes for different time windows that are used for calculating the explanatory variables. Both the window length and its endpoint relative to the time of trade are varied. Panel (a) presents the contour plot for period 1. Panel (b) presents it for period 3. The maximum likelihood window is three seconds long and stretches all the way until the time of trade. This window enters the probit regressions presented in table 2.3.

tive” intervals, i.e., the intervals that exhibit persistent price price declines. Alternatively, she strategically “takes liquidity” when there is occasional buy interest in the market and therefore becomes part of the “non-toxic” activity.

In summary, the probit results along with the graphs reveal that the large seller did take “price and time” into account when trading. This could, in addition to an imperfect prediction of future volume, explain why she did not always make the 9% volume target (see figure 2.6).

2.4 Price for immediacy

The large seller sold its 75,000 contracts at a substantial discount as evidenced by the timing of her trades (see, e.g., figure 2.1). A natural interpretation is that this is the price she paid for “demanding immediacy”. But, how much did she pay? And, was this a reasonable price to pay? These questions are explored in the remainder of this section.
Price paid for immediacy. Table 2.4 shows that the large seller paid between $56.4 and $117.3 million for immediacy depending on the choice of reference price. The average price she paid should be judged against a benchmark “unpressured price” or “fundamental value”. Natural choices for this benchmark price are the midquote that prevailed at 14:32 (the large seller’s first sell), 14:53 (the large seller’s last sell), and 16:15 (the market close). The table presents estimates for all three reference prices. The lowest cost estimate is obtained when judging against the price at the time of the large seller’s last trade: $56.4 million or, in relative terms, 135.3 basis points. If the general price after this last trade is interpreted as further recovery from the liquidity demanded, then the cost estimate jumps to $98.6 million or 234.1 basis points. The highest estimate is obtained when the benchmark price is the price that prevailed at the time of the large seller’s first trade: $117.3 million or 277.4 basis points. These cost estimates are not only substantial in and of themselves, they are also large relative to the size of the large seller. The table shows that, for example, these cost estimates range from about the level of her operating income that year or twice that level.

![Figure 2.8: A snapshot of trading by the large seller and others](image_url)

This figure depicts the aggressive sells by the large seller in a one minute window before the E-mini halt. The minute is part of period 1, the period when the E-mini and SPY price were cointegrated. The data is aggregated to 100-millisecond intervals. The size of the symbols in the graph scale with volume.
(a) Order flow toxicity (VPIN) in the Flash Crash period

![Graph showing order flow toxicity (VPIN) during the Flash Crash period.]

(b) Toxicity changes (ΔVPIN) and the large seller’s volume share

![Graph showing changes in VPIN and the large seller’s volume share.]

**Figure 2.9: Order flow toxicity and the large seller’s volume share.** This figure illustrates co-movement between order flow toxicity and the large seller’s volume share in the Flash Crash period. Panel (a) plots VPIN, the toxicity measure that was developed in Easley, López de Prado, and O’Hara (2012). The VPIN is calculated based on volume buckets. Its value is plotted on the natural time scale by anchoring each volume bucket to the time of the last trade. Panel (b) plots the change in VPIN along with the large seller’s volume share in each volume bucket.
2.4. PRICE FOR IMMEDIACY

Was the price that was paid reasonable? A simple calibration exercise reveals whether the price that the large seller paid for immediacy was reasonable. The model is inspired by one strand of the microstructure literature that is often referred to as “inventory models.” These models feature a risk-averse intermediary who might need to temporarily hold an inventory when intermediating between buyers and sellers. The price risk she runs over such positions requires her to charge those who demand immediacy a “price pressure.”

The calibration model is essentially the theoretical model proposed by Grossman and Miller (1988) with constant relative risk aversion instead of constant absolute risk aversion, CRRA instead of CARA. One important motivation for CARA utility in theoretical papers is that models can be solved analytically. CRRA utility, however, is the de facto choice for calibration (see, for example, Mehra and Prescott, 1985). The model itself is quite straightforward. A seller with an inelastic supply curve of 75,000 E-mini contracts trades with a representative CRRA intermediary who operates on a zero-profit condition (to capture a competitive market). After an interval of length $T$, the intermediary trades out of the position with buyers who provide him with an infinitely elastic demand curve. This makes him obtain the fundamental price when selling in the second stage. The price that clears the market in the first stage relative to the second-stage price captures the size of the “price pressure” that the seller had to pay. The deep parameters that drive the size of this pressure are: the intermediary’s relative risk aversion $\gamma$, her initial wealth level $w_0$, the length of the holding period $T$, and the price risk per unit of time $\sigma^2$. For simplicity, the intermediary’s discount rate is assumed to be one. This is a relatively innocent assumption as her inventory holding period is short (hours or, in the worst case, a couple of days).

Parameter choice. The volatility parameter $\sigma^2$ is estimated based on price changes in the period before the Flash Crash half hour, i.e., 9:30 through 14:30 on May 6. The size of the position, $q$, is the dollar value of 75,000 E-mini contracts at 14:32, the time of the large seller’s first sell. It amounts to $4.1$ billion. The remaining three parameters $w_0$, $T$, and $\gamma$ are allowed to vary in order to calibrate the observed price pressure. Three levels are chosen for intermediary wealth: $4.1$ billion, $78$ billion, and $200$ billion. $4.1$ billion is the minimum wealth level in the intermediation sector because CRRA utility does not admit negative wealth. $78$ billion was approximately the market capitalization of Goldman Sachs on May 6, 2010. Kirilenko et al. (2011) report a categorization which estimates that there were 16 high-frequency traders and more than 100 market intermediaries (not including cross-market arbitrageurs, or opportunistic traders, who might also serve as intermediary between the index futures market and index ETFs or between the index futures market and the market for individual stocks). Therefore, an estimate of $200$ billion for total sector market capitalization is relatively conservative (given that Goldman Sachs alone is $78$ billion).

Two values are considered for the length of the holding period. A “5-hour” period
### (a) Price pressure paid (bps)

<table>
<thead>
<tr>
<th>Time</th>
<th>Reference Price</th>
<th>Price Pressure Paid by the Large Seller</th>
<th>Loss Relative to Total Assets</th>
<th>Loss Relative to Operating Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>14:32</td>
<td>1127.63</td>
<td>117.3 (bps)</td>
<td>277.4 (%)</td>
<td>13.0 (%)</td>
</tr>
<tr>
<td>(large seller's first sell)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14:53</td>
<td>1111.38</td>
<td>56.4 (bps)</td>
<td>135.3 (%)</td>
<td>6.3 (%)</td>
</tr>
<tr>
<td>(large seller's last sell)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:15</td>
<td>1122.63</td>
<td>98.6 (bps)</td>
<td>234.1 (%)</td>
<td>11.0 (%)</td>
</tr>
<tr>
<td>(end of the day-trading session)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*a* Book value assets reported by large seller in her 2010Q1 quarterly report.

*b* Q1 operating income reported by large seller in her 2010Q1 quarterly report.

### (b) Calibrated price pressure (bps)

<table>
<thead>
<tr>
<th>Initial Wealth</th>
<th>5 Hours</th>
<th>3 Days</th>
<th>5 Hours</th>
<th>3 Days</th>
<th>5 Hours</th>
<th>3 Days</th>
</tr>
</thead>
<tbody>
<tr>
<td>$4.1 billion²</td>
<td>1.6</td>
<td>27.3</td>
<td>0.1</td>
<td>1.5</td>
<td>0.0</td>
<td>0.6</td>
</tr>
<tr>
<td>$78 billion³</td>
<td>6.3</td>
<td>106.5</td>
<td>0.4</td>
<td>5.9</td>
<td>0.2</td>
<td>2.3</td>
</tr>
<tr>
<td>$200 billion⁴</td>
<td>7.8</td>
<td>134.2</td>
<td>0.4</td>
<td>7.4</td>
<td>0.2</td>
<td>2.9</td>
</tr>
<tr>
<td>10⁵</td>
<td>15.5</td>
<td>263.1</td>
<td>0.8</td>
<td>14.8</td>
<td>0.3</td>
<td>5.8</td>
</tr>
<tr>
<td>5⁵</td>
<td>18.6</td>
<td>313.2</td>
<td>1.0</td>
<td>17.8</td>
<td>0.4</td>
<td>7.0</td>
</tr>
<tr>
<td>50⁵</td>
<td>76.7</td>
<td>1136.7</td>
<td>4.2</td>
<td>73.4</td>
<td>1.6</td>
<td>28.9</td>
</tr>
</tbody>
</table>

*a* Intermediary has just enough wealth to take the full position.

*b* Goldman Sachs market capitalization on May 6, 2010.

*c* Estimate from Hendershott and Menkveld (2013) based on NYSE specialist data.


*e* Estimate from Barsky et al. (1997) based on experiments.

*f* See chapter 21 of Cochrane (2005).

---

**Table 2.4: Price pressure paid by the large seller.** This table provides an analysis of the cost the large seller incurred for demanding “immediacy” on 75,000 E-mini contracts. This cost is referred to as the price pressure she paid. Panel (a) presents estimates of this cost by comparing the large seller’s average sell price to a reference (“fundamental”) price. Panel (b) calibrates a Grossman-Miller (1988) type model to the price paid for immediacy. It does so by documenting what “price pressure” a zero-profit representative intermediary with constant relative risk aversion utility would have charged. Three parameters vary: (i) the intermediary’s relative risk aversion, (ii) his wealth, and (iii) the inventory holding period.
assumes normal market conditions: According to CFTC and SEC (2010a) (page 2), the large seller had previously executed a similar size sell program: “On that occasion it took more than 5 hours for this large trader to execute the first 75,000 contracts of a large sell program.” The other holding period considered here is three days.

Finally, several relative risk-aversion coefficients for CRRA utility are used. Hendershott and Menkveld (2013) estimate that NYSE specialists for large stocks have, on average, a relative risk-aversion of 3.96. Mehra and Prescott (1985) use relative risk-aversion coefficients below 10. From experiments, Barsky et al. (1997) find that agents have an average relative risk-aversion of about 12. Using a classical asset pricing model, Cochrane (2005) (chapter 21) finds that a relative risk-aversion of more than 50 is necessary to explain stock returns, the risk-free rate, and consumption growth in 20th century U.S. markets.

**Calibration result.** Table 2.4 panel (b) presents the calibration result. It shows that the price pressure for “reasonable” values of the model’s parameters is an order of magnitude lower than the price pressure the large seller paid (the lower bound on the estimate is 234.1 basis points, see panel (a)). The calibrated price pressure matches the realized pressure only if all of the three following conditions hold: (i) the intermediary sector is relatively risk-averse ($\gamma \geq 5$), (ii) the aggregate wealth in the intermediation sector is thin (just enough to buy the $4.1$ billion position), and (iii) the expected holding period is long (three days). This area is indicated by a gray shading in panel (b).

### 2.5 Conclusion

This paper complements earlier empirical work on the Flash Crash with new evidence. It adds analysis based on a proprietary dataset provided by the large seller whose trading, reportedly, was a key part of the crash. By and large, the evidence reveals that her trading did not cause any of the steep price declines ahead of the E-mini halt in a direct manner. Her trades in this period however are followed with a 300 millisecond delay by strong net sells by others and, at that same time, a sharp price decline. The evidence further reveals that large seller strategy was driven by a volume ratio target, but she did not follow it blindly. For example, she timed her trades so as to benefit from transitory price increases. Finally, a simple Grossman-Miller type calibration exercise reveals that the “price pressure” she paid was disproportionately large.

The evidence should inform current regulatory debates. One reading of the new results is that the crash cannot be attributed to a single agent but really is the product of agent interaction. The unfortunate consequence is that the large seller overpaid for the immediacy it demanded. This, as such, is bad news for investors as they learn that there is, potentially, huge cost in unwinding a large position. This could lead to a higher required return or even
shying away from market participation altogether. A clear-cut recommendation does not emerge as it depends on the mechanism that drove the destructive interaction of agents in the market. Further study of candidate mechanisms is left for future theoretical work.

2.6 Appendix

2.6.A Data integrity

**Quote delays.** The CFTC-SEC report mentions substantial price quote delays in the Flash Crash period. Chapter III.3 of CFTC and SEC (2010a) documents that 1,665 NYSE-listed symbols were affected. The list did not include the NYSE-listed SPY. Nanex reports two instances of SPY quote delays in BATS, each lasting for about one second. Panel (a) of figure 2.10 illustrates one of them. These delays, however, seem short-lived and idiosyncratic relative to the multiple-minute arbitrage opportunity/cointegration breakdown documented in section 2.3.2. The CFTC-SEC report does not mention quote delays on the CME-traded E-mini contract, nor did we find any other source that reported of such delays.

**Trade report delays.** Trades might experience a time delay as exchanges are “required to report their trade activity within 90 seconds of execution time to the Consolidated Tape System (CTS)” (see appendix Q of SEC, 2001).

Before going through the data line by line, panel (b) of figure 2.10 plots the end-of-period bid-ask spread and the volume-weighted average trade price at a one-second frequency. The panel leads to the following observations. First, for all exchanges trade price appears to track the bid-ask spread up to the 90-second delay that is permitted. The only exception is CINC where, in the half hour of the Flash Crash, trade delay runs up to about three minutes. Second, quote and trade activity at CBOE and CHIC seems to shut down for five to ten minutes in the recovery stage of the Flash Crash. Third, the bid-ask spread is substantially wider after the halt, although it does widen in some exchange even before the halt (see, for example, CBOE).

**Line by line check.** This subsection performs a basic check on “data integrity” by going through the dataset line by line for each exchange. For quote records, the best quotes are compared to check if the best bid price equals (a “touch”) or exceeds (a “cross”) the best ask price. For trade records, a trade is counted as an “immediate match” if the trading price is the same as either the immediately preceding best ask or best bid quote. For SPY data, the integrity check also counts “delayed matches” for various delay lengths, a category discussed below. Table 2.5 summarizes the results. It distinguishes between the Flash Crash period (14:30-15:00) and the “normal” non-crash period (9:30-16:15).
Figure 2.10: Data feed. This figure illustrates the quality of the data feed by plotting the bid-ask spread and (average) trade price. Panel (a) does so for a couple of seconds of BATS data at a 25-millisecond granularity. It illustrates what is, most likely, a delay in the BATS quote feed. Panel (b) and (c) plot the spread and trade price for the Flash Crash half hour period at a one-second granularity. Panel (b) illustrates the E-mini data feed. Panel (c) illustrates the data feed of the eight exchanges that trade SPY.
Figure 2.10 continued…

(c) Trade and bid-ask spread in SPY
2.6. APPENDIX

(a) Flash Crash period (14:30-15:00)

<table>
<thead>
<tr>
<th>Exchange</th>
<th>#quotes</th>
<th>touches</th>
<th>crosses</th>
<th>#trades</th>
<th>immediate matches</th>
<th>-25ms' matches</th>
<th>-100ms' matches</th>
<th>-500ms' matches</th>
<th>-10s' matches</th>
<th>-90s' matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-Mini</td>
<td>316,000</td>
<td>1,363a</td>
<td>0</td>
<td>191,000</td>
<td>99.8%</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>BATS</td>
<td>269,000</td>
<td>0</td>
<td>0</td>
<td>50,000</td>
<td>65.4%</td>
<td>81.1%</td>
<td>95.9%</td>
<td>98.2%</td>
<td>99.3%</td>
<td>99.5%</td>
</tr>
<tr>
<td>BOST</td>
<td>100,000</td>
<td>1</td>
<td>8</td>
<td>22,000</td>
<td>48.3</td>
<td>66.7</td>
<td>91.3</td>
<td>96.7</td>
<td>98.8</td>
<td>99.2</td>
</tr>
<tr>
<td>CBOE</td>
<td>9,900</td>
<td>7</td>
<td>61</td>
<td>404</td>
<td>68.3</td>
<td>69.3</td>
<td>86.1</td>
<td>98.0</td>
<td>98.2</td>
<td>99.3</td>
</tr>
<tr>
<td>CHIC</td>
<td>28,000</td>
<td>0</td>
<td>0</td>
<td>547</td>
<td>55.4</td>
<td>72.6</td>
<td>90.3</td>
<td>97.4</td>
<td>99.8</td>
<td>99.8</td>
</tr>
<tr>
<td>CINC</td>
<td>49,000</td>
<td>6</td>
<td>10</td>
<td>1,800</td>
<td>28.2</td>
<td>30.3</td>
<td>41.0</td>
<td>57.5</td>
<td>64.8</td>
<td>83.0</td>
</tr>
<tr>
<td>ISEX</td>
<td>125,000</td>
<td>17</td>
<td>22</td>
<td>5,000</td>
<td>34.9</td>
<td>53.4</td>
<td>91.8</td>
<td>98.0</td>
<td>99.2</td>
<td>99.4</td>
</tr>
<tr>
<td>NQEX</td>
<td>171,000</td>
<td>0</td>
<td>0</td>
<td>144,000</td>
<td>71.7</td>
<td>81.7</td>
<td>90.5</td>
<td>93.8</td>
<td>96.6</td>
<td>98.8</td>
</tr>
<tr>
<td>PACF</td>
<td>229,000</td>
<td>0</td>
<td>0</td>
<td>48,000</td>
<td>71.3</td>
<td>83.5</td>
<td>94.9</td>
<td>97.7</td>
<td>99.3</td>
<td>99.6</td>
</tr>
<tr>
<td>all SPY</td>
<td>982,000</td>
<td>31</td>
<td>101</td>
<td>272,000</td>
<td>67.4</td>
<td>79.6</td>
<td>91.8</td>
<td>95.2</td>
<td>97.4</td>
<td>98.8</td>
</tr>
</tbody>
</table>

(b) May 6 trading hours, excl. Flash Crash period (9:30-14:30, 15:00-16:15)

<table>
<thead>
<tr>
<th>Exchange</th>
<th>#quotes</th>
<th>touches</th>
<th>crosses</th>
<th>#trades</th>
<th>immediate matches</th>
<th>-25ms' matches</th>
<th>-100ms' matches</th>
<th>-500ms' matches</th>
<th>-10s' matches</th>
<th>-90s' matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-Mini</td>
<td>1,980,000</td>
<td>0</td>
<td>0</td>
<td>839,000</td>
<td>99.9%</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>BATS</td>
<td>1,963,000</td>
<td>0</td>
<td>0</td>
<td>282,000</td>
<td>72.2%</td>
<td>86.3%</td>
<td>97.6%</td>
<td>99.2%</td>
<td>99.9%</td>
<td>99.9%</td>
</tr>
<tr>
<td>BOST</td>
<td>623,000</td>
<td>1</td>
<td>1</td>
<td>75,000</td>
<td>52.4</td>
<td>70.1</td>
<td>91.4</td>
<td>97.0</td>
<td>99.3</td>
<td>99.7</td>
</tr>
<tr>
<td>CBOE</td>
<td>58,000</td>
<td>0</td>
<td>0</td>
<td>655</td>
<td>50.4</td>
<td>53.9</td>
<td>76.6</td>
<td>91.3</td>
<td>97.7</td>
<td>98.5</td>
</tr>
<tr>
<td>CHIC</td>
<td>180,000</td>
<td>0</td>
<td>0</td>
<td>2,800</td>
<td>64.1</td>
<td>78.9</td>
<td>91.8</td>
<td>94.7</td>
<td>99.7</td>
<td>99.9</td>
</tr>
<tr>
<td>CINC</td>
<td>212,000</td>
<td>0</td>
<td>0</td>
<td>3,600</td>
<td>36.7</td>
<td>41.1</td>
<td>65.2</td>
<td>92.0</td>
<td>99.8</td>
<td>99.9</td>
</tr>
<tr>
<td>ISEX</td>
<td>735,000</td>
<td>12</td>
<td>3</td>
<td>12,000</td>
<td>48.4</td>
<td>69.0</td>
<td>96.2</td>
<td>99.2</td>
<td>99.7</td>
<td>99.9</td>
</tr>
<tr>
<td>NQEX</td>
<td>1,050,000</td>
<td>0</td>
<td>0</td>
<td>570,000</td>
<td>75.9</td>
<td>86.1</td>
<td>94.9</td>
<td>97.7</td>
<td>99.6</td>
<td>99.9</td>
</tr>
<tr>
<td>PACF</td>
<td>1,731,000</td>
<td>0</td>
<td>0</td>
<td>282,000</td>
<td>77.2</td>
<td>88.1</td>
<td>96.9</td>
<td>98.7</td>
<td>99.6</td>
<td>99.8</td>
</tr>
<tr>
<td>all SPY</td>
<td>6,551,000</td>
<td>13</td>
<td>4</td>
<td>1,232,000</td>
<td>73.3</td>
<td>85.1</td>
<td>95.4</td>
<td>98.0</td>
<td>99.4</td>
<td>99.9</td>
</tr>
</tbody>
</table>

a All these touching quotes occurred during the 5-second trading halt (14:45:28–14:45:33).

Table 2.5: Data integrity check. This table presents some basic data integrity checks. Panel (a) reports statistics for the Flash Crash half hour period. Panel (b) reports the same statistics for the other trading hours on that day. The left-most columns focus on quotes. They report (i) the number of quote updates, (ii) the number of times the best bid equals the best ask (“touch”), and (iii) the number of times it exceeds it (“cross”). The right-most columns focus on trades. They report the number of trades as well as how many of them can be matched with a best bid or ask quote that preceded it. The matching is done based on either the bid and ask quote immediately preceding it or all quotes in a time interval just ahead of the trade report. The interval lengths considered are 25 milliseconds, 100 milliseconds, 500 milliseconds, 10 seconds, or 90 seconds. This exercise is motivated by the 90-second time delay that is allowed for exchanges when distributing their trade reports. There is no such time delay for distributing quotes.
E-mini. E-mini contracts traded only on CME Globex. The sequencing of quote and trade records seems to be reliable. The 2.3 million quote records in the day-trading hours do not contain crossing quotes and all touching quotes occurred during the five-second trading halt from 14:45:28 to 14:45:33. In the Flash Crash period, more than 99.8% of the trade records can be matched with either the best bid or the best ask quote that immediately preceded it. If instead of a simple count, mismatches are weighted by volume, the matching rate also exceeds 99.8%. In the non-crash period, the matching rate is above 99.9%.

SPY. SPY traded on eight exchanges: BATS, BOST, CBOE, CHIC, CINC, ISEX, NQES, and PACF. Table 2.5 shows that occurrences of the best bid touching or crossing the best ask are few. Added up, these events occur in 0.01% of all quote events during the Flash Crash period and they occur even less frequently in the non-crash period.

The table further documents that trade records matching is much more problematic. In the Flash Crash period, only 67.4% of the trade records can be matched with immediately preceding quote prices. This low matching rate, however, is a more general characteristic of the market as also in the non-crash period the rate is low: 73.3%. The 90 second slack exchanges have on reporting trades suggests that trade prices should be matched with a window of price quotes of up to 90 seconds. Matching rates improve substantially when allowing for delay. For example, they become 91.8% and 95.4% for a window length of 100 milliseconds and 98.8% and 99.9% for a window length of 90 seconds. This general pattern is fairly consistent across exchanges except for CINC. In that case even the 90-second matching rate during the Flash Crash period remains as low as 83.0% (consistent with figure 2.10).

Data used for analysis. To sum up, the data quality check leads to the following two observations. First, price quotes seem largely reliable as intra-market touches and crosses are few and timestamps do largely line up with the timestamps on trades given the 90 second delay allowed for trade reporting (the only exception is CHIC during the Flash Crash period, but CHIC generates less than one percent of all trades; see table 2.5). All analysis based on price quotes therefore uses all available data without applying any filters.

Second, the timestamp on trades is an unreliable indicator of execution time as exchanges are allowed to delay trades by up to 90 seconds. The analysis most troubled by this is the table 2.2 analysis that relates price changes to net flow based on a VAR/VECM analysis. Net flow is obtained by signing trades according to the side of initiation; a trade obtains a positive sign if it was a buyer who hit an ask quote and a negative sign if a seller hit a bid quote. For each exchange, the signing is done by comparing trade price with the bid and ask quotes that precede it. In the Flash Crash period, on average, 99.8% of E-mini trades and 67.4% of SPY trades can be signed this way. Tick rule is used for all other trades (see, e.g., Lee and Ready, 1991). More than 90% of the mismatches are, arguably,
2.6. APPENDIX

the result of a delay less than 100 milliseconds (see table 2.5). It is for this reason that the VAR/VECM analysis aggregates the data over 100 millisecond intervals. This seems to be a fair trade-off between (i) a low frequency to reduce the risk of erroneous sequencing and (ii) a high frequency to identify the interrelationship between all variables in the system. To further improve data quality, trades from CINC (many delays beyond 90 seconds), CBOE and CHIC (activity disappears for several minutes) are removed. In total, these trades make up one percent of all trades in the Flash Crash period. The result of all other analysis is not sensitive to trade timestamp inaccuracies in the order of 100 milliseconds.

2.6.B Breakpoints estimation

Breakpoint estimation for time series models has been extensively studied in, for example, Andrews (1993), Andrews, Lee, and Ploberger (1996), Bai (1997), and Bai and Perron (1998). These papers consider a coefficient change at some point in time as a structural break. The econometric task is to estimate the coefficients in the subsamples before and after the break along with the time at which the break occurs. The consistency and efficiency of the estimators critically depend on the time series properties of all variables in the candidate subsamples. For example, the procedure requires the variables to be stationary throughout. To the best of our knowledge, breakpoint estimation that involves cointegration breaks has not been explored. This appendix proposes an approach that accommodates cointegration breaks. The new approach extrapolates from the existing approaches for stationary series.

Maximum-likelihood is used to estimate the time points of the cointegration break and recovery for the E-mini and SPY price series (this general pattern is visible in figure 2.2 and discussed on page 15). Denote the E-mini mid-quote price by \( p^E(t) \) and the SPY mid-quote price by \( p^S(t) \). The sample period runs from 14:30 to 15:00 with timestamps that are accurate up to 25 milliseconds. The time index \( t \) therefore runs from 0 to 72,000. The price differential series is defined as

\[
(2.1) \quad z(t) = \left[ \ln p^E(t) - \ln p^E(0) \right] - \left[ \ln p^S(t) - \ln p^S(0) \right].
\]

Due to the log transformation and the subtraction of the \( t = 0 \) price, \( z(t) \) can be interpreted as the cumulative return differential between E-mini and SPY.

The following three-step procedure is used to estimate the breakpoints \( T_1 \) and \( T_2 \):

1. For all possible integer pairs \((T_1, T_2) \in [0, 72000] \times [0, 72000]\) perform an augmented Dickey-Fuller test (with both drift and time trend) on \( z(t) \) for the three subperiods defined by: \( \tau_1 = [0, T_1), \tau_2 = [T_1, T_2), \text{ and } \tau_3 = [T_2, 72000]. \) If the pair \((T_1, T_2)\) is such that the null of a unit root is rejected for \( \tau_1 \) and \( \tau_3 \), but not for \( \tau_2 \) then proceed to step two. If not, discard the pair.
2. Use maximum likelihood to estimate a time series model for each subperiod. The models proposed for the three periods are:

\[ \phi_i(L) z(t) = \varepsilon(t), \text{ for } t \in \tau_i \text{ and } i \in \{1, 2, 3\}, \]

where \( \{\varepsilon(t)\} \) is (normally distributed) white noise. All roots of the polynomials \( \phi_1(\cdot) \) and \( \phi_3(\cdot) \) are outside the unit circle. The polynomial \( \phi_2(\cdot) \) has exactly one unit root and all its other roots are outside the unit circle.

3. The pair \((T_1, T_2)\) that yields the highest overall likelihood (obtained by multiplying the three likelihoods) is chosen to be the breakpoint pair.

Three implementation issues merit discussion. First, the number of lags in \( \phi_i \) is unknown for each subperiod \( \tau_i \). The AR polynomials are truncated at 40 lags for all three subperiods, in both the augmented Dickey-Fuller test (step 1) and the maximum likelihood estimation (step 2). Second, to ensure that each regression has sufficiently many observations, the length of \( \tau_i \) is set to be at least 10 seconds (4,000 observations). Third, the augmented Dickey-Fuller tests in step 1 use a significance level of 5%.
Middlemen Interaction

How modern financial market is affected by middlemen multiplicity

Middlemen multiplicity is a new phenomenon in modern financial markets: a multiple of market making high-frequency traders stand in the process of financial asset transaction, intermediating between fundamental investors. This chapter takes a theoretical approach to analyze the consequences on market quality. It is presumed that two frictions, arrival asynchronicity and information asymmetry, stand in the way of efficient asset reallocation from a distressed low-valuation seller to high-valuation buyers. Informed, well-connected middlemen alleviate these frictions by matching investors across markets and by generating market activity from which the uninformed seller can learn. However, the learning is imperfect as the seller cannot distinguish price pressure associated with inter-middlemen trades from a fundamental value drop. When middlemen’s reselling opportunities are small, additional middlemen might hurt allocative efficiency and, in particular, add to the cost of the distressed seller. The analysis speaks to recent disruptions in securities markets (e.g., the Flash Crash; see chapter 2).
This chapter is based on Menkveld and Yueshen (2014b). An early version was largely developed from my M.Phil. thesis, which owes many credits to the thesis committee members Albert J. Menkveld, Peter Gautier, and Enrique Schroth. Apart from the later collaboration with Albert J. Menkveld, the discussions with and comments from Jos van Bommel, Giovanni Cespa, Carole Comerton-Forde, Sarah Draus, Thierry Foucault, Terrence Hendershot, Bengt Holmström, Albert S. Kyle, Katya Malinova, Gideon Saar, and Haoxiang Zhu are gratefully acknowledged. The current version has acquired immense benefit from the participants at 40th EFA meeting (Cambridge, U.K.), FIRS 2013 (Dubrovnik), HFT conference (Paris), 16th SGF Zurich, Advances in AT and HFT (UCL), U of Bath, Carlos III, 10th International Paris Finance Meeting, 2012 Market Microstructure (Institut Louis Bachelier), U of Lugano, Central Bank Workshop (Bundesbank), U of Toulouse, U of Piraeus, the 5th Erasmus Liquidity Conference, U of Luxembourg, Skye Workshop 2012 (U of Aberdeen, U of Bath, and Swansea U), Cass Business School, Warwick U, Finance Down Under 2012 (U Melbourne), and Erasmus U.

3.1 Introduction

Over the past two decades, securities markets around the world have gradually moved to electronic trading. In 2010, the SEC (2010) reviewed the equity market structure. Its report highlights two salient features. First, trading became highly segmented; a single name is traded across multiple lit and dark venues. Second, high-frequency trader (HFT), has arrived at the market. HFTs reportedly participated in more than half of U.S. equity trades by 2010. The SEC characterizes them as professional traders who 1) use extraordinarily high-speed computer programs to generate, route, and execute orders and 2) maintain very short time frames for establishing and liquidating positions (SEC, 2010, p.45). One interpretation is that HFTs function as middlemen, a perspective maintained throughout this paper.\footnote{1 In the taxonomy of HFT proposed by the SEC, this perspective fits their categories a and b (market making and arbitrage). The remaining categories c and d that involve exploitation of structural vulnerabilities, order anticipation strategies, and momentum ignition strategies are beyond the scope of this paper. Hagströmer and Nordén (2013) shows that the market-making HFTs constitute the majority of HFT activity in NASDAQ-OMX Stockholm.}

The role of middlemen in a securities markets has been extensively studied. Early models focus on the optimizing behavior of market makers in the presence of a very reduced-form investors demand for liquidity.\footnote{2 See, e.g., Ho and Stoll (1983), Glosten and Milgrom (1985), Easley and O’Hara (1987), Grossman and Miller (1988), Biais (1993), Lyons (1997), and, more recently, Cespa and Foucault (2013).} In parallel, a literature grew on the optimal execution problem of a “distressed” investor who needs to execute a large order quickly.\footnote{3 See, e.g., Keim and Madhavan (1995), Chan and Lakonishok (1995), Bertsimas and Lo (1998), and Almgren and Chriss (2001), Brunnermeier and Pedersen (2005), Engle and Ferstenberg (2007), Gârleanu and Pedersen (2013), and Obizhaeva and Wang (2013).} A more
3.1. INTRODUCTION

recent literature sets out to integrate both sides, market makers and investors, into a trade economy to study welfare. This paper adds to this literature by modeling the interaction of investors and middlemen in a segmented market environment where middlemen, equipped with super computing technology, are both well-connected and well-informed. Notably, the model allows middlemen interaction to feed back to the distressed investor’s optimal execution problem. This characterization fits the modern electronic trading environment.

The paper asks several questions and proposes a model to analyze them. Why and how middlemen trade with one another? How do these inter-middlemen trades affect the trading of others, e.g., a distressed seller who seeks liquidity? Is market quality helped by adding middlemen? The model identifies and articulates channels through which middlemen interaction affects market quality, judged by allocative efficiency (i.e., welfare). It is shown that whether adding more middleman is good critically depends on how easily they can resell their inventory to investors, or in the terminology of this paper, their reselling opportunities. In particular, when reselling opportunities are small, additional middlemen might hurt welfare.

Each middleman’s RO is modeled jointly by two random variables. A first (Bernoulli) draw determines whether the RO exists or not: Either the middleman has RO or he does not. Heterogeneous realizations of these random variables prompt middlemen to trade with one another. Conditional on that the middleman has RO, a second random draw determines whether the reselling attempt (to an end-user investor) succeeds. Across the market, the (average) success rate, termed naturally as “RO size”, is summarized by a single parameter which reflects the general (exogenous) market condition, e.g. fundamental investors’ willingness to participate. It turns out that this RO size parameter has important implications on market quality.

Inter-middlemen trades occur when the middleman-owner of the asset has no RO while another middleman has one. The middleman-owner, who otherwise suffers cost of carry, sells the asset to the other middleman, but only at a discount referred to as price pressure. This discount is required since the receiving middleman only has a reselling opportunity. It might still unfortunately fail in which case the receiving middleman has to hold the asset and pay the cost of carry. In contrast, trades between a middleman and an investor (end-user, void of the cost of carry) do not require such price pressure. Notably, a middleman’s expected cost of carry is endogenous in the model: It increases in the RO size as the middleman becomes less likely to get stuck with the asset.

Based on the middlemen characterization, the model studies asset reallocation across investors: from a distressed low-valuation seller, who has suffered an exogenous shock, to high-valuation buyers. Two frictions stand in the way of the efficient transfer: investor arrival asynchronicity (Grossman and Miller, 1988; Duffie, 2010) and information asym-
CHAPTER 3. MIDDLEMEN INTERACTION

The trade economy is modeled after Lagos and Rocheteau (2007) and Lagos, Rocheteau, and Weill (2011). In an ex ante period, the seller has the option to trade with middlemen (who are always around). In an interim period middlemen optimally trade with one another or with a buyer (if reselling succeeds). Although these two periods, extrapolated from and consistent with Grossman and Miller (1988), can conceptually repeat indefinitely, in order to close the model, the repetition is squeezed into one ex post period where the seller meets the sufficiently many informed buyers. The novelty of the model is to allow the seller to learn about the fundamental value from market activity in the interim period. This will help her reduce the adverse-selection cost she incurs in the ex post period.

The model identifies three channels through which middlemen affect market quality. On the positive side, middlemen alleviate the two frictions 1) by helping the early seller find buyers; and 2) by generating market activity that is informative to the uninformed seller as she learns about the asset’s true value. However, middlemen interaction also hinders efficient allocation because, 3) when there is market activity, the seller’s learning is imperfect as she cannot tell whether the trade is price-pressured (inter-middlemen trade) or not (middleman-buyer trade).

The model’s main result is that market quality is helped by adding middlemen only when reselling opportunities are large. This result can be understood from the channels through which middlemen interaction affects efficient reallocation. If the ROs are small, then additional middleman do not add a lot to the connection to find buyers in the interim stage (channel 1). Similarly, the appearance of market activity is not affected much, either (channel 2). It does however magnify the price pressure (if present) in the interim stage and thus make the seller's inference problem more severe (channel 3). When ROs are small enough, the third channel dominates and adding middlemen both harms the seller and reduces allocative efficiency. Section 3.4 finds an exact decomposition of the effect of an additional middleman on welfare and provides a detailed discussion.

Further, the model predicts that the seller always “overloads” the middlemen sector: Welfare could be improved if she would reduce her supply when trading with middlemen in the ex ante stage. This is because by offloading the asset to middlemen early, the seller avoids being adversely selected by more informed, late arriving buyers. From a social planner’s perspective, the seller’s adverse-selection cost offsets the buyer’s equivalent gain. That is, the seller has a higher benefit than the social planner of selling the asset to middlemen early. Hence, the quantity that the seller and the middlemen trade is too large relative to the first-best as the seller does not internalize the positive externality she exerts on the buyer when adversely selected.

The model sheds new light on recent disruptions in electronic markets. For example, on May 6, 2010, U.S. equity indices declined by 5 to 6 percent, and then recovered, all in 20 minutes: this event is known as the Flash Crash. CFTC and SEC (2010a) observe that (p.1-3) against a “backdrop of unusually high volatility and thinning liquidity, a large
3.1. INTRODUCTION

fundamental trader (a mutual fund complex) initiated a sell program to sell a total of 75,000 E-Mini contracts (valued at approximately $4.1 billion) as a hedge to an existing equity position.” In their view, this triggered a “liquidity crisis” in which, during the one minute of the extreme price drop of 4%, “the same positions were rapidly passed back and forth” by the high-frequency traders (inter-middlemen trades). This observation is corroborated in Kirilenko et al. (2011) who study disaggregated data on the E-mini market. The sudden evaporation of the buy-side market depth (less than 1% of the level in the morning; see p.3 CFTC and SEC, 2010a) indicates a extremely small RO size. Then, consistent with the inference problem emphasized by the current model, “the large trader responded to the increased volume by increasing the rate at which it was feeding the orders into the market, even though orders that it already sent to the market were arguably not yet fully absorbed by fundamental buyers or cross-market arbitrageurs. In fact, especially in times of significant volatility, high trading volume is not necessarily a reliable indicator of market liquidity” (p.3, CFTC and SEC, 2010a). Menkveld and Yueshen (2014a) show that the large seller paid a disproportionately large cost for trading immediately during the Flash Crash. The Flash Crash does not seem to be a singular event. It has since been followed by disruptions in other markets, like the “mini Flash Crash” in German DAX index future (August 26, 2011) and many others.¹

This current paper relates to the “hot-potato” literature on dealership markets; see, among others, Lyons (1997), Naik, Neuberger, and Viswanathan (1999), and Viswanathan and Wang (2004). The key innovation of the current paper is the motivation of inter-middlemen trades. Instead of risk-sharing as in “hot-potato” trades, it is the heterogeneous ROs that drive middlemen to trade with one another. As the number of middlemen increases, the inter-middlemen price pressure identified in the current paper does not change as it is only determined by middlemen’s RO size. However, the risk-sharing “hot-potato” price pressure would gradually diminish with more middlemen, because in aggregate they have larger risk-bearing capacity.² In addition, that middlemen interaction can both help and hurt other investors via different channels is the new insight brought about by the current model.

Finally, this paper adds to the high-frequency trading (HFT) literature and, in particular, market making HFTs. Hagströmer and Nordén (2013) analyze NASDAQ-OMX Stockholm data and find that most HFT quote traffic and volume originates market making types. Baron, Brogaard, and Kirilenko (2013) analyze HFT activity in futures markets and document that HFT types are persistent through time; HFTs seem to specialize. Menkveld and

¹ “Mini flash crashes: A dozen a day”, CNNMoney. March 20, 2013.
² In a dynamic extension, conceptually, RO driven trades might aggravate price pressure as middlemen realize that it is harder and harder to find buyers, while risk-sharing driven trades reduce price pressure. In addition, repeated trades among middlemen are much harder to explain under pure risk-sharing, as, unlike risk aversion, ROs might change very frequently.
Zoican (2013) show, both theoretically and empirically, that HFT market makers and its complement group, opportunistic HFTs, are affected in different ways to a latency drop. In particular, the adverse-selection cost for market makers is increased. Menkveld (2013) documents how an HFT market maker actively trades across two platforms to manage its inventory. A comprehensive review of the recent HFT literature can be found in Jones (2013).

The rest of the paper is structured as follows. Section 3.2 sets up the model and equilibrium is analyzed in section 3.3. Welfare analysis is the focus of section 3.4. Section 3.5 discusses the model assumptions as well as some extensions. Section 3.6 concludes. The appendix contains a notation summary (section 3.7.A) and proofs (section 3.7.B).

### 3.2 Model primitives

As a convention, random variables are denoted in upper case, while their realizations and other deterministic variables are in lower case. Greek letters denote probabilities.

The following economic environment is consistent with Lagos, Rocheteau, and Weill (2011).

**Goods and the asset.** There is a numéraire good and a special good for consumption. A perfectly divisible asset is traded in the market. Each unit of the asset pays off one unit of the special good at the end of the time (see timeline below). Only the asset is traded, but not the special good. Denote the amount of the numéraire good by $c$ and the amount of the asset by $a$, which is also the amount of the special good upon consumption.

**Agents.** There are infinite investors who initially have the same preference $u(c, a) = c + Z \cdot a$, where $Z \geq 0$ is the random value of the special good, uniformly distributed on $[0, 2]$ (with unity mean).

One of these investors suffers a negative shock and becomes the seller, $S$, whose preference changes to

$$u^S(c, a) = \begin{cases} 
  c + Z \cdot a, & \text{if } a < a^* \\
  c + Z \cdot (a - k(a - a^*)), & \text{if } a \geq a^*
\end{cases}$$

where $-k(a - a^*)$, defined for $a > a^*$, is $S$’s (negative) private value for the special good and it is three-times differentiable with $k(0) = k_a(0) = 0$ and for $a \geq 0$, $0 \leq k_a \leq 1$, $k_{aa} \geq 0$, and $k_{aaa} \leq 0$.\(^1\) $S$ carries an endowment of $a_0 > a^*(> -\infty)$ units of the asset (that is, $S$ would

---

\(^1\) Equivalently, one can define $v(a) := a - k(a - a^*)$ such that $v(a)$ is defined on $[a^*, \infty)$, three-times differentiable, and concavity increasing in $a$ with $v(a^*) = a^*$ and $v_a(a^*) = 1$. $v_{aaa} = -k_{aaa} \geq 0$ follows necessarily if $v(a)$ has decreasing absolute risk-aversion. Note that assuming $k_a \leq 1$ ensures that $S$’s marginal utility in consuming the special good is always positive.
not want to sell more than \( a_0 - a^* \) units of the asset). Normalize \( a^* = 0 \).

All other investors are referred to as buyers, \( \mathbf{B} \), with preference \( u^\mathbf{B}(c, a) = c + Z \cdot a \) unchanged.

There are \( m \) (\( \geq 1 \), integer) homogeneous middlemen. The middlemen are labeled by \( \mathbf{M}_i \) with \( i \in \{1, \ldots, m\} \). Their preference is characterized by \( u^\mathbf{M}(c, a) = c + 0 \cdot a = c \), in order to capture the notion that middlemen are not the ultimate consumers of the special good (see also Duffie, Gârleanu, and Pedersen (2005), Lagos and Rocheteau (2009), and others).

A middleman will be referred to as “it” (because these middlemen are typically computers in the context). The seller will be referred to as “she” and a buyer will be referred to as “he”.

Instead of focusing on search and bargaining as in Lagos, Rocheteau, and Weill (2011), the model addresses agents’ asynchronous arrival in the sense of Grossman and Miller (1988), or “investment inattention” as in Duffie (2010).

**Timeline.** There are three periods: ex ante, interim, and ex post. The Grossman and Miller asynchronicity is embedded in the ex ante and the interim periods. In the ex ante period, \( \mathbf{S} \) and all \( \mathbf{M} \) are in the market but there is no \( \mathbf{B} \). In the interim period, all \( \mathbf{M} \) are in the market and there *might* be a \( \mathbf{B} \). The above two periods can, conceptually, repeat many times but in order to simplify the analysis, the model is wrapped up in the ex post period where \( \mathbf{S} \) meets \( \mathbf{B} \) directly.\(^1\) The asset then pays off the special good and consumption takes place. Figure 3.1 summarizes the timing.

In the ex ante period, \( \mathbf{S} \) finds an \( \mathbf{M} \), for example, the fastest one, and trades with it. Label this \( \mathbf{M} \) by \( \mathbf{M}_1 \). In the interim period, \( \mathbf{M}_1 \) has three options to resell and it can choose only one of them: 1) to look for an early \( \mathbf{B} \) in the market—the *common* channel, 2) to look for an early \( \mathbf{B} \) in its *private* channel, or 3) to look for another \( \mathbf{M} \) in the market to trade with. (Short-selling is not allowed.) If \( \mathbf{M}_1 \) is indifferent between multiple channels, it randomly chooses one with equal probability weight on each. In the ex post period, \( \mathbf{S} \) meets (a representative of) \( \mathbf{B} \) and they trade.

**Trading.** All trades are conducted with discriminatory pricing (chapter 5.3 of Vives, 2008) in order to capture the cost of information asymmetry.\(^2\) The trade initiator first offers an inverse supply function \( p(s) \), which specifies the marginal price \( p \) for the \( s \)-th unit of the

---

\(^1\) The consequence of allowing \( \mathbf{M} \) to also present in the ex post period is discussed later in section 3.5.

\(^2\) The trading scheme is to reflect the fact that the trade initiator has to signal his/her interest to trade and therefore is subject to the potential adverse-selection cost if the counterparty is more informed. The scheme is consistent with, for example, trading in a limit order market where limit orders are subject to the pick off risk by informed traders. See also Biais, Martimort, and Rochet (2000, 2013), Wang and Zender (2002), Viswanathan and Wang (2004), and Back and Baruch (2013) among many others for discriminatory pricing.
Figure 3.1: Timing and information. This figure summarizes the timing of the model. There are three periods: ex ante, interim, and ex post. The agents’ decisions and their information are shown in each period.

asset. The counterparty then decides the quantity to trade. The sell side initiates the trade; that is, S in ex ante, M1 in interim, and S in ex post.

Reselling opportunity (RO). An RO measures the likelihood for an M to find and trade with a B. Each Mi has access to the common RO denoted by B0 in the market and to a private RO denoted by Bi in its private channel. Each Bj is a Bernoulli random variable with success probability Φj for j ∈ {0, 1, ..., m} such that \( P(B_j = 1 | Φ_j) = Φ_j \). That is, Mi can find a B via RO j with probability Φj. At the beginning of the interim period, Nature decides whether each of these ROs exists. While some M have ROs, some might not: For j ∈ {0, 1, ..., m},

\[
Φ_j = \begin{cases} 
\phi, & \text{with probability } θ \text{ (RO exists)}; \\
0, & \text{with probability } 1 - θ \text{ (RO does not exist)}. 
\end{cases}
\]

Here, θ is interpreted as the probability for an RO to exist and φ, conditional on its existence, the size of the RO. A constraint on φ and θ is imposed in inequality (3.22) to avoid triviality (see appendix 3.7.D). The random variables Z, Φ0, Φ1, ..., and Φm are all independent of each other.

Information. Distributions of the random variables are common knowledge to all agents. In the ex ante period, no random variable has realized. In the interim period, each Mi observes z, the common φ0, and its own private φi (but not the others’). In the ex post period, S observes the interim trades, if any, in the market (but not those in M’s private channels). Finally, Bs always observe z.
3.3. EQUILIBRIUM

Model discussion. The discussion of the model assumptions, together with how they affect the model implications, is deferred to section 3.5.

3.3 Equilibrium

Section 3.3.1 first solves the optimal strategy of $M$ in the interim period. $S$ ’s ex post and ex ante optimization problems are then solved backwardly in sections 3.3.2 and 3.3.3, respectively.

3.3.1 $M$ in the interim period

In the interim period, there is an $M_1$ who holds $a_1 > 0$ units of the asset (bought from $S$ in the ex ante period) and wants to resell. There are three channels to look for a counterparty to trade with: 1) the common channel (reselling in the market), 2) the private channel, or 3) reselling to another $M$ (if $m \geq 2$). There is no information asymmetry about $Z$ between $M_1$ and its potential trading counterparty (either a $B$ or another $M$). Therefore, the best $M_1$ can do is to post a flat supply schedule at the targeted counterparty’s reservation price for the asset.

If $M_1$ looks for $B$ in the market, with probability $\phi_0$ it finds a $B$ and then sells $a_1$ units of the asset to him at price $\rho = z$ ($B$ also has reservation value $z$). The expected (interim) profit is $\phi_0 za_1$. Similarly, if $M_1$ looks for $B$ in its private channel, it gets expected profit of $\phi_1 za_1$. $M_1$ is indifferent between the common and the private channels if both are available, i.e. if $\phi_0 = \phi_1 = \phi$.

$M_1$ can always find another $M$, say $M_2$, if $m \geq 2$, as middlemen are always in the market. $M_2$ would like to buy from $M_1$ if and only if $M_2$ has non-zero probability to resell: either $\Phi_0 = \phi$ or $\Phi_2 = \phi$. Knowing this, when $M_1$ finds $M_2$, $M_1$ would like to offer $M_2$ at price $\phi z$, the reservation price for $M_2$. Note that the trading price is discounted: $\phi z < z$, and the difference

$$ (3.1) \quad (1 - \phi)z $$

is the price pressure that $M_2$ requires to compensate its expected cost of carry (in case $M_2$ cannot find a $B$ to resell). Note that the price pressure is endogenously determined by $\phi$, the size of ROs. $M_1$ can find such an $M_2$ as soon as at least one of $\{\Phi_0, \Phi_2, \ldots, \Phi_m\}$ is positive, i.e. with probability $\mathbb{P}(\sum_{i \neq 1} \Phi_i > 0 | \Phi_0)$.

Comparing these three channels, the following lemma concludes this subsection.

**Lemma 3.1 (M1 ’s strategy in the interim)**. Let $M_1$ holds $a_1 (> 0)$ units of the asset and observes the realizations $\phi_0$ and $\phi_1$. Then 1) if $\phi_0 = \phi$ and $\phi_1 = 0$, it looks for $B$ in the market (the common channel); 2) if $\phi_0 = 0$ and $\phi_1 = \phi$, it looks for $B$ in its private
3) if $\phi_0 = \phi_1 = \phi$, it looks for $B$ either in the market or in its private channel with equal probability; and 4) if $\phi_0 = \phi_1 = 0$, it looks for another $M$ in the market.

Following the lemma, the unconditional probability for $M_1$ to find and trade with a $B$ in the common channel is

$$\alpha := \left[ P(\Phi_0 = \phi, \Phi_1 = 0) + \frac{1}{2} P(\Phi_0 = \Phi_1 = \phi) \right] \phi$$

and the unconditional probability for $M_1$ to find and trade with another $M$ is

$$\beta(m) := P(\Phi_0 = \Phi_1 = 0) \cdot P\left( \sum_{i \geq 2} \Phi_i > 0 \right) = (1 - \theta)^2 \left( 1 - (1 - \theta)^{m-1} \right),$$

where $m$ is the total number of middlemen. Naturally, if $m = 1$, $\beta(m) = 0$.

### 3.3.2 S in the ex post period

Depending on $m$ (the number of $M$) and the realization of ROs, $S$ learns differently from the market activity from the interim. Table 3.1 summarizes her possible inference. There are three possible scenarios.

- Fully revealing, "FR". $S$ observes a very high interim trading price $P_1 > 2\phi$, so high that it is not possible to be associated with MM price pressure (the maximum of pressured trading price is $\sup \phi Z = 2\phi$.) In this case, $S$ perfectly infers that the true asset (common) value is $Z|_{P_1 > 2\phi} = P_1$.

- Partial learning, "PL". If the interim trading price is $P_1 \leq 2\phi$, $S$ cannot tell whether the trade is an “MB” trade (no price pressure, hence $Z = P_1$) or an “MM” trade (with price pressure, hence $Z = P_1 / \phi$). These are the two shaded lines shown in table 3.1.

- No learning, "NL". If there is no interim trade, $S$ can learn nothing. This is the case when $M_1$ goes to its private channel or when $M_1$ fails to find a counterparty in the market.

$S$ ’s optimal ex post supply is analyzed below separately for each of these three scenarios.
3.3. EQUILIBRIUM

<table>
<thead>
<tr>
<th>( \Phi_0 )</th>
<th>( B_0 )</th>
<th>( \Phi_1 )</th>
<th>( \sum_{t&gt;2} \Phi_t )</th>
<th>M1 looking for</th>
<th>interim trade</th>
<th>trading price, ( P_1 )</th>
<th>probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \phi )</td>
<td>1</td>
<td></td>
<td></td>
<td>B in market (found)</td>
<td>MB</td>
<td>( = Z \in (2\phi, 2] )</td>
<td>( \alpha \cdot (1 - \phi) )</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>&gt; 0</td>
<td></td>
<td>another M (found)</td>
<td>MM</td>
<td>( = Z \in [0, 2\phi] )</td>
<td>( \alpha \phi )</td>
</tr>
<tr>
<td>( \phi )</td>
<td>0</td>
<td></td>
<td></td>
<td>B in market (failed)</td>
<td>no interim trade</td>
<td>( = \phi Z \in [0, 2\phi] )</td>
<td>( \beta )</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td>B in private</td>
<td>(in the market)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>another M (failed)</td>
<td>no interim trade</td>
<td>(in the market)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1: S’s inference from the interim market activity. This table summarizes what inference about the asset value S might draw from the interim market activity. If the observed trading price \( P_1 \) is very high (> 2\phi), so high that there cannot be price pressure, the interim trade must be between an M and a B. If \( P_1 \) is relatively low (\( \leq 2\phi \)), it might either reflect a true asset value traded between M and B (\( P_1 = Z \)) or an inter-middlemen (MM) trade with price pressure (\( P_1 = \phi Z \)), as highlighted in the two shaded lines. Finally, when there is no activity from the interim period, S learns nothing. The last column summarizes the probabilities. The unconditional probabilities of MB and of MM trades are denoted respectively by \( \alpha \) and \( \beta \), as defined in equations (3.2) and (3.3).

3.3.2.1 Fully revealing

S perfectly infers that \( Z = p_1 \) (\( \geq 2\phi \)) in this scenario. There is no information asymmetry between S and B in the ex post period. Therefore, knowing B’s linear preference, S will offer to sell all her asset holding at \( p = z \), a flat supply curve, to B. B will be just willing to buy all that S offers. The unconditional expected utility for S is

\[
\bar{u}_{FR}^S(a_2) = \mathbb{E}[u_{FR}^S(a_2, Z) | Z > 2\phi] = a_2 \mathbb{E}[Z | Z > 2\phi] = (1 + \phi)a_2,
\]

where \( a_2 \) is her ex post position of the asset. Note that S no longer suffers the negative private value, \( k(\cdot) \), because she is able to sell off everything. Nor does she suffer any adverse-selection cost, i.e. the information rent that transfers to B in this case is zero.

3.3.2.2 Partial learning

S observes an interim trading price \( P_1 \leq 2\phi \) in this scenario. The conditional (on \( P_1 \leq 2\phi \)) probability of the observed trade being MM is

\[
\hat{\beta}(m) = \frac{\beta(m)}{\beta(m) + \alpha \phi},
\]
where $\alpha$ and $\beta(m)$ are the unconditional probability of an MB trade and of an MM trade, respectively, as defined in equations (3.2) and (3.3). (Appendix 3.7.C gives in detail how to compute this conditional probability.) The conditional distribution of the common value, $Z$, becomes

$$
\hat{Z}(p_1) := Z|p_1 = p_1 \leq \phi = \begin{cases} 
p_1/\phi, & \text{with probability $\hat{\beta}(m)$;} 
p_1, & \text{with probability $1 - \hat{\beta}(m)$.}
\end{cases}
$$

Intuitively, as $S$ does not know whether the observed trade is price pressured or not, her posterior belief about $Z$ is either high ($p_1/\phi$) or low ($p_1$). The following proposition states precisely S’s inference problem.

**Proposition 3.1 (S’s inference problem from the interim trade).** Suppose there is an interim trade at price $p_1 \leq \phi$. $S$ has an inference problem about the asset value from the observed trade: $\text{var}(Z|p_1 = p_1 \leq \phi) > 0$ if and only if $m \geq 2$. The inference problem persists even if there are infinitely many $M$.

If $m = 1$, $\hat{\beta} = \beta = 0$ and $\hat{Z} = p_1$ with probability 1; that is, when there is only one $M$, no MM trade is possible and this partial learning scenario degenerates to the fully revealing scenario. On the other hand, the MB probability is $\alpha > 0$ and it is not affected by the number of $M$ (equation 3.2). Therefore, even if there were infinitely many middlemen, making MM trade unconditionally extremely likely, $S$ still has the inference problem.

Consider $S$’s optimization problem. Because there are only two possible true posterior values for $Z$, $p_1/\phi$ (high) and $p_1$ (low), it suffices to solve the optimal (cumulative) supply quantities, $s_h$ (at high) and $s_l$ (at low), at these two prices:

$$
\max_{0 \leq s_l \leq s_h \leq a_2} \hat{\beta} \cdot \left[ \frac{p_1}{\phi} a_2 - \left( \frac{p_1}{\phi} - p_1 \right) s_l - \frac{p_1}{\phi} k(a_2 - s_h) \right] + (1 - \hat{\beta}) \left[ p_1 a_2 - p_1 k(a_2 - s_l) \right],
$$

where $a_2$ is the amount of the asset that $S$ still holds in the ex post period. Clearly, selling anything at other prices would be suboptimal. Note that the term $-(p_1/\phi - p_1)s_l$ represents $S$’s adverse-selection cost: When the true price is $p_1/\phi$, $S$ sells $s_l$ units of the asset at $p_1$, too low compared to the true value (winner’s curse). On the other hand, $-p_1k(a_2 - s_h)/\phi$ and $-p_1k(a_2 - s_l)$ represent her negative private value as she is unable to sell off all her undesired holdings. In choosing her supply, $S$ essentially trades off between the adverse-selection cost and her negative private value of the asset. The solution, presented by an inverse supply function, $p(s)$, is given by the following lemma.
3.3. EQUILIBRIUM

Figure 3.2: S’s ex post supply with partial learning. The blue, solid lines illustrate S’s supply in the ex post period after she observes a trade at price \( p_1 \leq 2\phi \). Whether or not S supplies anything at all at the observed trading price \( p_1 \) (the low price) depends on \( \hat{k}_a \), the severity of adverse-selection. If \( \hat{k}_a \) exceeds her holding \( k_a(a_2) \), she supplies nothing at the low price (the dashed line in panel (b)); if it does not, she sells the remainder \( a_2 - \hat{a}(m) \) at the low price (panel (a)). The shaded area in panel (a) represents the adverse selection cost of S when the true price is high: \( Z = p_1/\phi \).

Lemma 3.2 (S’s ex post supply after partial learning). Suppose S observes an interim trade at price \( p_1 \leq 2\phi \) and holds \( a_2 \) units of the asset in the ex post period. S then posts an ex post supply of

\[
p(s) = \begin{cases} 
    p_1, & \text{if } 0 < s < a_2 - k_a^{-1}(\min\{\hat{k}_a, k_a(a_2)\}) \\
    p_1/\phi, & \text{if } s \geq a_2 - k_a^{-1}(\min\{\hat{k}_a, k_a(a_2)\}) 
\end{cases}
\]

where \( \hat{k}_a := \beta \cdot (1 - \phi)/(\alpha \phi^2) \) reflects S’s (marginal) adverse selection cost.

Figure 3.2 plots the supply schedule. Panels (a) and (b) depict the cases of \( k_a(a_2) > \hat{k}_a \) and of \( k_a(a_2) \leq \hat{k}_a \), respectively. At \( p_1/\phi \) (the high price), S always sells everything she holds because there is no adverse-selection at that price \( (\mathbb{P}(Z > p_1/\phi | P_1 \leq 2\phi, p_1) = 0) \). At \( p_1 \) (the low price), S trades off between her negative private value for the asset and the expected adverse selection cost, such that, in expectation, the marginal adverse selection cost (\( \hat{k}_a \), see the proof of lemma 3.2 in the appendix) equates her marginal negative private value. The larger is \( \hat{k}_a \), the more severe is the adverse-selection cost that S is subject to. In particular, when \( \hat{k}_a \) is too large, the “lemon” market breaks down at the low price (Akerlof, 1970), as shown in panel (b).
As the number of middlemen increases, 3.2

The first conditional expectation,

\[ \mathbb{E}[P_1|P_1 = \phi Z] \]

which only arises following an MB trade. Similarly, S ’s expected adverse-selection cost, i.e. the equivalent transfer to B, is

\[ \tilde{\tau}_{PL}(a_2) := \hat{\beta} \cdot (1 - \phi)(a_2 - k_a^{-1}(\min\{\hat{k}_a, k_a(a_2)\})) \]

which only arises following an MM trade. Of course, one can verify that the following accounting identity holds:

\[ \hat{\beta} \mathbb{E}[Z] + (1 - \hat{\beta})\phi \mathbb{E}[Z] a_2 = \hat{a}_{PL}(a_2) + \tilde{\tau}_{PL}(a_2) + \tilde{\tau}_{PL}(a_2), \]

where the left-hand side expression is the expected value (conditional on the existence of some interim activity) of a_2 units of the asset, and the right-hand side is its decomposition (recall that \( \mathbb{E}[Z] = 1 \).

1 The first conditional expectation, \( \mathbb{E}[P_1|P_1 = \phi Z] \), corresponds to the case of observing a price-pressured \( P_1 \) that follows an MM trade. The second conditional expectation, \( \mathbb{E}[P_1|P_1 = Z, Z \leq 2\phi] \), corresponds to the case of observing an unpressured \( P_1 \) that follows an MB trade.
Figure 3.3: S’s ex post supply in when there is no learning. This figure sketches S’s optimal supply function if there is no learning for her. The supply is strictly increasing. If the true value of Z is at z, the shaded area represents the seller’s adverse-selection cost and the unsold quantity, \( a_2 - s(z) \), reflects her negative private value \( k(a_2 - s(z)) \). Note that at \( s = 0 \), \( p = 2/(k_0(a_2) + 1) > 1 \) because \( k_0(a) \) is bounded from above by 1.

3.3.2.3 No learning

Without learning anything, S holds the prior belief that Z is uniform on \([0, 2]\). With \( a_2 \) units of the asset, she tries to solve an optimal supply schedule \( s(\cdot) \) that maximizes her expected utility:

\[
\max_{s(\cdot) \geq 0} \mathbb{E} \left[ Z \cdot a_2 - Z \cdot k(a_2 - s(Z)) - \int_0^Z s(p) dp \right].
\]  

(3.10)

It can be seen that S trades off her negative private value \( k(a_2 - s(Z)) \) and the adverse selection cost \( \int_0^Z s(p) dp \). The solution of this calculus of variation problem is given (in the form of an inverse supply function) by the following lemma and it is graphed in figure 3.3.

Lemma 3.3 (S’s ex post supply when there is no learning). Suppose there is no learning from the interim period and S holds \( a_2 \) units of the asset. Then S posts a supply schedule of \( p(s) = 2/(k_0(a_2 - s) + 1) \), where \( 0 < s \leq a_2 \).

As before, the following expressions for S’s expected utility and costs will prove useful in analyzing the ex ante period game (section 3.3.3) as well as in the welfare analysis (section 3.4). Using equations (3.10) and lemma 3.3, S’s utility in this scenario can be
evaluated as
\[
\begin{align*}
\bar{u}_{NL}^S(a_2, Z) &= \begin{cases} 
    Z \cdot (a_2 - k(a_2)), & \text{if } 0 \leq Z \leq \bar{z}(a_2) \\
    \bar{z}(a_2) a_2 - Z k(k_a^{-1}(2/Z - 1)) + \int_{\bar{z}(a_2)}^{Z} k_a^{-1}(2/p - 1) \, dp, & \text{if } \bar{z}(a_2) < Z \leq 2^\cdot
\end{cases}
\end{align*}
\]
where \( \bar{z}(a_2) := 2/(k_a(a_2) + 1) \in (1, 2] \). Taking the unconditional expectation, one can get S’s expected utility for holding \( a_2 \) units of the asset if she learns nothing from the interim period:

\[
(3.11) \quad \bar{u}_{NL}^S(a_2) := \int_{\bar{z}(a_2)}^{Z} Z \cdot (a_2 - k(a_2)) \, f(Z) \, dZ + \int_{\bar{z}(a_2)}^{Z} \left[ \bar{z}(a_2) a_2 - Z k(k_a^{-1}(2/Z - 1)) + \int_{\bar{z}(a_2)}^{Z} k_a^{-1}(2/p - 1) \, dp \right] \, f(Z) \, dZ.
\]

In this case, S’s expected (negative) private value after no learning as

\[
(3.12) \quad \bar{k}_{NL}(a_2) := \int_{\bar{z}(a_2)}^{Z} Z k(a_2) \, f(Z) \, dZ + \int_{\bar{z}(a_2)}^{Z} Z k(k_a^{-1}(2/Z - 1)) \, f(Z) \, dZ
\]

where \( f(Z) = 1/2 \) is the density of \( Z \). Similarly, define S’s expected adverse-selection cost after no learning as

\[
(3.13) \quad \bar{\eta}_{NL}(a_2) := \mathbb{E} \left[ \int_{\bar{z}(a_2)}^{Z} s(p) \, dp \right] = \int_{\bar{z}(a_2)}^{Z} \int_{\bar{z}(a_2)}^{Z} \left[ a_2 - k_a^{-1}(2/Z - 1) \right] \, dp \, f(Z) \, dZ.
\]

Note that \( \bar{\eta}_{NL}(a_2) \) is the equivalent gain of B. One can verify the following accounting identity with some further calculation:

\[
(3.14) \quad \mathbb{E} Z \cdot a_2 = a_2 = \bar{u}_{NL}^S(a_2) + \bar{k}_{NL}(a_2) + \bar{\eta}_{NL}(a_2).
\]

Closed-form expressions for the above integrals do not exist in general, unless some additional assumptions on \( k(a) \) (e.g., \( k(a) = a - \ln(1 + a) \)) are made. However, it is possible to characterize the shapes of the above gains and losses functions. In particular, carefully computing the first- and second-order derivatives of \( \bar{u}_{NL}^S \) confirms that this expected utility is indeed concavely increasing in the asset holding:

\[
(3.15) \quad \frac{\partial}{\partial a_2} \bar{u}_{NL}^S(a_2) = \frac{1}{1 + k_a(a_2)} > 0; \quad \text{and} \quad \frac{\partial^2}{\partial a_2^2} \bar{u}_{NL}^S(a_2) = -\frac{k_{aa}(a_2)}{(1 + k_a(a_2))^2} < 0.
\]

This concavity will be used below to ensure the existence of the equilibrium in the ex ante period. Similarly, the transfer is also monotone increasing; the larger is S’s ex-post holding, the higher is B’s information gain:

\[
(3.16) \quad \frac{\partial}{\partial a_2} \bar{\eta}_{NL}(a_2) = \frac{k_a(a_2)^2}{(1 + k_a(a_2))^2} > 0.
\]
3.3. EQUILIBRIUM

3.3.3 S in the ex ante period

There is no information asymmetry between S and M in the ex ante period. S therefore knows the reservation price of M1. From table 3.1, this reservation price can be computed as

\[
p_0(m) = \alpha \cdot (1 - \phi)\mathbb{E}[Z | Z > 2\phi] + \alpha \phi \mathbb{E}[Z | Z \leq 2\phi] + \alpha \mathbb{E}Z + \beta(m)\mathbb{E}[\phi Z] = 2\alpha + \beta(m)\phi
\]

(3.17)

where the first two components are the expected revenue from reselling to B in the market, the third component is the expected revenue from reselling privately, and the last component is the expected revenue from reselling to another M.\(^1\)

The difference between the asset’s unconditional expected value and \(p_0\), \(\mathbb{E}Z - p_0 = 1 - p_0\), is the price discount that is just necessary to compensate M’s expected cost of carry. Note that \((1 - p_0)\) is monotone decreasing in \(\phi\), i.e. the larger is the RO size, the lower is M1’s (expected) cost of carry; equivalently, \((1 - p_0)\) is also the probability of failing to resell in the interim period (recall that \(\mathbb{E}Z\) is normalized to unity). Intuitively, when ROs are small, M1 finds it very hard to resell to B and the cost of holding the asset becomes large in expectation as M do not enjoy the special good.

As the optimal price is solved, it remains for S to choose the quantity \(q_0\) to supply at this price, given her initial shock size \(a_0\):

\[
\bar{u}_S = \max_{0 \leq q_0 \leq a_0} \begin{cases} p_0q_0 & \text{selling early} \\ + \alpha \cdot (1 - \phi)\bar{u}_{\text{FR}}(a_0 - q_0) + (\alpha \phi + \beta)\bar{u}_{\text{PL}}(a_0 - q_0) + (1 - \alpha - \beta)\bar{u}_{\text{NL}}(a_0 - q_0) & \text{selling late, fully revealing} \\ \text{selling late, partial learning} & \text{selling late, no learning} \end{cases}
\]

where \(\bar{u}_{\text{FR}}, \bar{u}_{\text{PL}},\) and \(\bar{u}_{\text{NL}}\) are given in equations (3.4), (3.7), and (3.11). The trade-off that S faces is selling the asset early at price \(p_0 < \mathbb{E}Z = 1\) (the discount is to compensate M’s expected cost of carry) and selling late in which case she is subject to her negative private value and/or adverse-selection cost.

The analysis and the solution to this optimization problem is deferred to appendix 3.7.D, where a technical condition is assumed (inequality 3.22) to ensure interior solution. Instead, the following result is presented here to facilitate the welfare analysis later in section 3.4.

**Proposition 3.3 (S’s ex ante loading on M).** The more middlemen there are, the more S loads them in the ex ante period. Mathematically, \(q_0(m)\) increases in \(m\).

---

\(^1\) Note that the unconditional probabilities of successfully reselling in the market and in the private channel are the same \(\alpha\).
CHAPTER 3. MIDDLEMEN INTERACTION

The intuition is as follows. There are three effects of an additional $M$ on $S$ ’s expected utility: 1) $p_0(m)$ is increased as there are more RO via $M$; 2) the probability of MM trade is increased, hurting $S$ (corollary of proposition 3.2); and 3) the probability of having no interim market activity is reduced. The first two effects, interestingly, offset each other: The seller will sell to $M$ more of the asset in the ex ante period to just exactly offset the the negative effect of her worsened inference problem in the interim; see the proof of proposition 3.3 for details. The third effect, hence also the net effect, improves $S$ ’s expected marginal utility because, unconditionally, there are more market activity to be observed and she can learn from them.

3.4 Welfare

This section analyzes social welfare, measured as the total ex ante expected utility of all agents in this economy, $w = u^S + u^M + u^B$. $S$ ’s ex ante expected utility can be evaluated from equation (3.20) using the solution derived in appendix 3.7.D. The ex ante expected utility for $M$ is zero, because their reservation price is known to $S$ in the ex ante period. $B$ ’s ex ante expected utility arises from adversely selecting $S$ in the ex post period, when $S$ reinforces the MM price pressure (proposition 3.2) or when she does not learn anything (no market activity). After some simplification, the resulting welfare expression is

\[ w = a_0 - (1 - p_0(\beta))q_0 - \alpha \phi^2 \hat{k}(\beta) - (1 - \alpha - \beta)\tilde{k}_{NL}(a_0 - q_0), \]  

(3.18)

where $q_0$ is the loading from $S$ to $M$ ex ante, $\hat{k}(\beta)$ is a short-hand notation for $k(k_a^{-1}(\hat{k}_a(\beta)))$, and $\tilde{k}_{NL}(\cdot)$ is defined in equation (3.12). The first term, $a_0$, is the first-best welfare (as the expected common value is $EZ = 1$). In the above expression, it is emphasized that $p_0$ and $\hat{k}$ are functions of $\beta$, which is in turn a monotone function in $m$, the number of $M$ (equation 3.3).

Equation (3.18) helps understand the three channels of potential welfare loss, identified as the three minus terms in equation (3.18). First, $(1 - p_0)q_0$ is the expected amount of the asset that are not able to be resold by $M$, when the RO realizations are bad. The second and the third part are $S$ ’s (negative) private value due to her non-zero holding after an MB trade and after learning nothing, respectively. In these two cases, $S$ does not sell everything in the ex post because she does not want to be adversely selected. Her holding, from a social planner’s perspective, is inefficient because of her negative private value $k(\cdot)$. Note that the second welfare loss part, $\alpha \phi^2 \hat{k}$, is due precisely to the inference problem created by MM trades (proposition 3.1) and it strictly increases with $m$ as more $M$ elevate $S$ ’s adverse-selection risk (corollary of proposition 3.2).
Proposition 3.4 (Overloading). In equilibrium, \( S \) always “overloads” \( M \) in the ex ante period in the sense that if \( S \) were to reduce supply by a small amount, welfare could be improved.

The intuition of this proposition is as follows. Consider a welfare-maximizing social planner who dictates how much of the asset that \( S \) should sell to \( M \) in the ex ante period at price \( p_0 \). Compared to the preference of \( S \), this social planner also takes into account the potential gain of \( B \) in the ex post trade. Such gains by \( B \) come at the adverse-selection cost of \( S \). As \( S \) “ignores” the potential positive externality (benefiting \( B \)) of her ex post selling, she sells more asset up front than the social planner—she overloads the middlemen.

With the help of propositions 3.3 and 3.4, an exact decomposition of the marginal effect of an additional \( M \) on welfare can be constructed. Recall that \( \beta \), the unconditional probability of MM trade, is monotone increasing in \( m \), the number of \( M \), and, as can be seen from the welfare expression (3.18), \( \beta \) is also the only channel through which an additional \( M \) affects welfare. Thus, it suffices to check the following partial derivative:

\[
\frac{\partial w}{\partial \beta} = \left( \frac{\partial p_0}{\partial \beta} q_0 + \tilde{k}_{NL}(a_0 - q_0) - \alpha \phi^2 \frac{\partial \hat{k}}{\partial \beta} \right) + \frac{\partial w}{\partial q_0} \frac{\partial q_0}{\partial \beta}.
\]

The first three components are direct effects.

- \((\partial p_0/\partial \beta)q_0\) reflects effect of the added RO brought by the additional \( M \). As the transfer from \( S \) to \( B \) becomes more likely (more RO), this effect is positive on welfare.

- \(\tilde{k}_{NL}(a_0 - q_0)\) corresponds to the effect of the reduced probability of \( S \) learning nothing. The additional \( M \) makes MM trade more likely, hence, unconditionally, more market activity is visible to \( S \). This effect is also positive on welfare.

- \(\alpha \phi^2 \partial \hat{k}/\partial \beta\) reflects the consequence of increased probability of MM trade. More MM trade makes \( S \)’s inference problem, conditional on a trade, worse, imposing larger adverse-selection cost to her in expectation. Hence, to avoid the increased adverse-selection cost, \( S \) reduces her ex post supply, and the inefficient holding becomes larger. This effect reduces welfare.

There is a fourth, an indirect, effect.

- The first part, \(\partial w/\partial q_0\), is the “overloading” effect, which is always negative because \( S \) overlooks her positive externality to \( B \) ex post (proposition 3.4). The second part, \(\partial q_0/\partial \beta\), is always positive because the more \( M \) there are, the more \( S \) sells to \( M \) ex ante (proposition 3.3). This increased overloading effect is, therefore, negative.
The overall effect of these four components critically depend on the RO size, \( \phi \).

**Proposition 3.5 (Additional M on welfare).** An additional M always monotonically improves welfare only if the reselling opportunity size is large.

When RO size becomes large, in the extreme \( \phi \to 1 \), each M is guaranteed reselling and there is no price pressure in an MM trade (equation 3.1). As such, S’s inference problem and adverse selection cost vanishes. In this case, additional M simply enlarge aggregate RO and hence improve welfare. On the other hand, when RO size is very small, the contribution of an additional M to the aggregate RO is limited, and yet the M worsens S’s learning. Further, the MM price pressure increases because of the elevated cost of carry. In fact, the third effect in the welfare decomposition (equation 3.19) dominates when \( \phi \) is small. Welfare, therefore, is reduced by additional M in this case.

Remark. Proposition 3.5 only characterizes the shape of welfare function (with respect to the number of middlemen) very limitedly. In particular, it does not state a sufficient condition for monotone-increasing welfare function: It only states a necessary condition (see the proof for more details). Further characterization on the shape of the welfare function is possible with additional restrictions on the cost function \( k(\cdot) \).

Figure 3.4 numerically illustrates the change of welfare with respect to the number of M using the parametrization of \( a_0 = 100, \theta = 0.8 \), and \( k(a) = a - \ln(1 + a) \) (for \( a \geq 0 \)). The threshold \( \bar{\phi} \) is approximately 0.902. It can be seen from the contour plot in panel (a) that along a horizontal line at some \( \phi > \bar{\phi} \), welfare monotonically increases from left to right. This, however, is not the case for \( \phi < \bar{\phi} \). Panel (b) zooms into the shaded area in panel (a), by showing the welfare changes along three levels of \( \phi \in \{0.88, 0.9, 0.92\} \). Clearly, along the curve of \( \phi = 0.92 \), welfare improves monotonically with additional M. However, along \( \phi = 0.88 \), welfare monotonically decreases. (Though the pattern is not very obvious, along \( \phi = 0.9 \), welfare first decreases and then increases.)

### 3.5 Discussion

This section discusses some of the model’s assumptions and argue that the key results of the paper remain robust even if these assumptions are relaxed.

**No M in the ex post period.** The model assumes that M do not participate in the ex post period. This is motivated by that investment inattention or arrival asynchronicity is of less concern in the long-run than in the very short-run. To the extent that all agents participate in the long-run (the ex post period), the M essentially become B and their cost of carry would disappear because they can always resell to true investors.
3.5. DISCUSSION

Figure 3.4: The effect of additional M on welfare—a numerical illustration. This figure numerically illustrates the net effect of additional M on welfare, measured as percentage losses relative to the first-best allocation. The numerical procedure assumes the following parametrization: $a_0 = 1$, $\theta = 0.8$, and $k(a) = a - \ln(1 + a)$ (for $a \geq 0$). Panel (a) shows the contour plot of welfare on the support of $(\beta(m), \phi)$. RO size increases along the vertical axis, while the number of middlemen increases along the horizontal axis (recall that $\beta(m)$ is monotonically increasing in $m$ and is bounded from above by $\beta(\infty) = (1 - \theta)^2$). The horizontal dashed line shows the threshold of $\bar{\phi}$, which is approximately 0.902 under the parametrization. Panel (b) zooms into the light-blue shaded area of panel (a). The vertical axis shows welfare and the horizontal axis represents the number of middlemen via $\beta(m)$. Three welfare curves are plotted, corresponding to the RO sizes of $\phi \in \{0.88, 0.9, 0.92\}$.

When M do participate in the ex post trade, by, for example, buying from S and then reselling to B, they profit from adversely selecting S whenever S fails to learn Z perfectly. This way, M would have an incentive to prevent S from learning Z from the interim market activity. This motivation would only strengthen the key result of the paper that more M could impair investors’ learning ability. Note that S’s optimal behavior is not affected by such M in the ex post period, because the adverse-selection remains the same regardless of by M or by B.

Information structure. The information structure about the asset value $z$ can be thought of as an extreme case of Jovanovic and Menkveld (2012). The information asymmetry between S and B is motivated by that the late agents (B) can always observe more than the
early one \((S)\) does.

It is assumed that \(M\) also perfectly observe the fundamental value in the interim period. This information edge is motivated by recent advancement in trading technology which enables the super computers to read and parse news much faster than humans (see also Foucault, Hombert, and Roşu (2013) and Biais, Foucault, and Moinas (2013) among others). The extreme view that \(M\) perfectly observes the true value can be relaxed. \(S\) ’s learning channel from interim market activity persists as long as MB trades contain information about the true asset value.

**S trading with only one M in the ex ante period.** In the current version of the model, \(S\) is only allowed to trade with one \(M\) in the ex ante period (the fastest \(M\) gets all). When \(S\) is allowed to trade with multiple \(M\), these multiple middlemen all holds non-zero inventory position to the interim period and all seek to resell, to \(B\) in the market, to private \(B\), or to one another. The interim market activity can no longer be described easily by a single trade as is in the current model (table 3.1), but rather by the number of trades, the trading volume of each trade, and the trading prices of each trade. This expanded information set appears to partition the state finer, but \(S\) ’s inference problem is not solved. Specifically, in an interim period with \(B\) in the market, all \(M\) trade with these \(B\) and all trading prices will be at the fundamental value. On the other hand, in an interim period with no \(B\) in the market, the market activity corresponds to MM trades and the associated prices are pressured. \(S\) still cannot distinguish MB trades from MM trades.

### 3.6 Conclusion

This paper studies the effect of middlemen interaction on market quality in an economy where a low private value seller looks for dispersed and asynchronously arriving high private value buyers. In addition, standard information asymmetry stands in the way of efficient asset allocation: the seller is subject to the adverse selection by the buyers, who arrive late and are information-driven.

Middlemen help the seller find more buyers in the short-run by bringing more reselling opportunities to the market and, further, their market activity, either trading with the buyers (MB trades) or trading with one another (MM trades), enables the seller to learn (maybe imperfectly) about the asset value. These two channels improve efficient allocation of the asset, hence enhancing market quality. However, as the number of middlemen increases, the seller’s inference problem (conditional on a trade) gets worse. This is because, compared to an MB trade, an MM trade carries price pressure in order to compensate the asset holder middleman for its cost of carry. Unable to tell whether the observed trade is price pressured,
the seller’s learning ability is impaired. The increased total transaction cost for the seller hinders the efficient allocation of the asset. This channel reduces welfare.

The overall effect of additional middlemen on market quality critically depends on the size of the reselling opportunity, i.e. how likely each middleman can find a buyer in the short-run. When the reselling opportunity size is large, more middleman improves welfare and when the reselling opportunity size is small, the negative effect dominates and more middleman can reduce market quality.

The analysis speaks to the electronic trading environment where, instead of a single middleman connecting buyers and sellers, multiple of them (in particular, high-frequency traders) operate in the market making business. The interaction of these multiple middlemen has proved to be of great market impact as witnessed in the Flash Crash and many other similar market disruptions.

3.7 Appendix

3.7.A Notation summary

- $a$, units of the asset. It also equates the amount of the special good upon consumption.
- $B$, the (representative) buyer.
- $B_j$, indicator of whether a buyer ($B$) arrives in the interim in channel $j$. It is a Bernoulli random variable with realization 1 for arrival and 0 for no arrival. For $j = 0$, it refers to the $B$ arriving in the market, i.e. the common channel. For $j \in \{1, \ldots, m\}$, it refers to the $B$ arriving in $M_j$’s private channel.
- $c$, units of the numéraire good.
- $k(a)$, S’s (negative) private value for the special good.
- $\hat{k}_a$, S’s marginal adverse-selection cost after observing an interim trading price $P_1 \leq 2\phi$.
- $\bar{k}_{NL}(a_2)$, S’s expected private value of holding $a_2$ units of the asset when trading with $B$ in the ex post period if there is no learning from the interim for her.
- $m$, number of middlemen.
- $M_i$, middleman $i$, $i \in \{1, \ldots, m\}$. In particular, $M_1$ refers to the asset owning middlemen in the interim period and $M_2$ refers to $M_1$’s counterparty if $M_1$ finds one.
- $p_t$, price in period $t$; $t = 0$ for the ex ante period and $t = 1$ for the interim period.
- $q_0$, quantity of the asset supplied by $S$ in the ex ante period.
- RO, acronym for reselling opportunity.
- $S$, the seller.
- $u(c, a)$, utility function of an agent who consumes $c$ units of the numéraire good and $a$ units of the special good. Superscript, one of “S”, “M”, and “B”, indicates the agent type.
The first-order condition with respect to \( q \). Fix a realized \( \varphi \). If \( \phi = 0 \) and \( \varphi_1 = \varphi \), then looking for \( B \) in \( M_1 \)'s private channel gives \( \phi z a_1 > \mathbb{P}(\sum_{i \geq 2} \Phi_i > 0) \phi z a_1 > 0 \), strictly dominating the other two options. When both \( \phi_0 = \varphi_1 = \varphi \), \( M_1 \) can get the same expected profit from either of the two channels. It randomly picks one with equal probability. Finally, if both \( \phi_0 \) and \( \varphi_1 \) are zero, only by trading with another \( M \) might \( M_1 \) get positive expected profit.

\[ \beta(m) \] the unconditional probability of an MM trade.

\( \Phi_i \) the random variable for the RO of channel \( i \), where \( i = 0 \) for the common channel in the market and \( i \geq 1 \) for \( M_i \)'s private channel. It realizes to be either \( \phi > 0 \), with probability \( \theta \) or \( 0 \) (no RO), such that \( \mathbb{P}(B_i = 1|\Phi_i) = \Phi_i \).

\( \varphi \), the size of \( M \)'s RO; that is, with probability \( \phi \), an \( M \) can find and trade with a \( B \) in the interim period if it has such an RO.

\( \theta \), the probability that an RO realizes to be “high”, that is, \( \mathbb{P}(\Phi_j = \phi) = \theta \) and \( \mathbb{P}(\Phi_j = 0) = 1 - \theta \).

3.7.B Proofs

Lemma 3.1

\textit{Proof.} If \( \phi_0 = \phi \) and \( \varphi_1 = 0 \), \( M_1 \) has expected profit of \( \phi z a_1 \) in the market, strictly dominates the other two options. If \( \phi_0 = 0 \) and \( \varphi_1 = \varphi \), then looking for \( B \) in \( M_1 \)'s private channel gives \( \phi z a_1 > \mathbb{P}(\sum_{i \geq 2} \Phi_i > 0) \phi z a_1 > 0 \), strictly dominating the other two options. When both \( \phi_0 = \varphi_1 = \varphi \), \( M_1 \) can get the same expected profit from either of the two channels. It randomly picks one with equal probability. Finally, if both \( \phi_0 \) and \( \varphi_1 \) are zero, only by trading with another \( M \) might \( M_1 \) get positive expected profit. \( \square \)

Proposition 3.1

\textit{Proof.} Fix a realized \( p_1 \leq 2 \phi \). Then following the conditional distribution in equation (3.6), \( \text{var}(Z|P_1 = p_2 \leq 2 \phi) = \hat{\beta}(m) (1 - \hat{\beta}(m)) p_1^2 (1/\phi - 1)^2 \). Note from equation (3.5) that \( \hat{\beta} \) is strictly bounded from above by \( (1 - \theta)^2 < 1 \) and reaches the lower bound 0 if and only if \( m = 1 \). Hence, whenever \( m \geq 2 \), the inference problem exists for \( S \), even if \( m \to \infty \). \( \square \)

Lemma 3.2

\textit{Proof.} The first-order condition with respect to \( s_h \) is \( \hat{\beta}(m) \frac{\partial}{\partial \varphi} k_a(a_2 - s_h) = 0 \), which has unique solution at \( s_h = a_2 \), i.e. selling everything at the high price. The first-order condition with respect to \( s_i \) is \( -\hat{\beta}(m)(p_1/\phi - p_1) + (1 - \hat{\beta})p_1 k_a(a_2 - s_i) = 0 \), or \( k_a(a_2 - s_i) = \hat{k}_a := \beta \cdot (1 - \phi)/(\alpha \phi^2) \), where \( \hat{k}_a \) can be interpreted as the substitution ratio (scaled by the inverse
hazard rate of $\hat{Z}$) of the marginal unit of the asset. The larger $\hat{k}_a$ is, the more severely is $S$ adversely selected.

Note that $k_a(a_2 - s_l)$ is bounded from above by $k_a(a_2)$ because the supply cannot be negative (even if $S$ wanted to “buy”, there would be no counterparty to trade with). Therefore, if $\hat{k}_a > k_a(a_2)$, the solution is cornered at $s_l = 0$, and if not, the interior solution is $s_l = a_2 - k_a^{-1}(\hat{k}_a)$. More compactly, $s_l = a_2 - k_a^{-1}(\min\{\hat{k}_a, k_a(a_2)\})$. Inverting $s_l$ and $s_l$, one can get the optimal inverse supply function as stated in lemma 3.2. The second-order conditions hold by the assumption on the convexity of $k(\cdot)$.

**Proposition 3.2 and its corollary**

**Proof.** Lemma 3.2 gives $S$’s optimal ex post supply function upon seeing an interim price $p_1 \leq 2\phi$. If the true state is MM, implying $Z = p_1 / \phi$ at the high price, $S$ will have sold $s_l$ units of the asset at the low price to $B$ ex post. This means that $S$ reinforces the MM trading price pressure. If the true state is MB, implying $Z = p_1$ at the low price, $S$ will only have sold $s_l \leq a_2$ units of the asset to $B$ ex post. (The inequality is strict if $k_a(a_2) < \hat{k}_a$.) This means that there are $a_2 - s_l$ units of the asset are still inefficiently held by $S$. Both $S$ and social welfare suffer from such inefficiency.

The corollary immediately follows as the (marginal) adverse-selection cost, $\hat{k}_a$ is strictly increasing in $m$ (see the expression from lemma 3.2) and $S$’s negative private value $p_1 k(a_2 - s_l) = p_1 k(\min\{k_a(a_2), k_a^{-1}(\hat{k}_a)\})$ following an MB trade is weakly increasing in $\hat{k}_a$.

**Lemma 3.3**

**Proof.** Rewrite the optimization problem (3.10) as

$$\max_{s(\cdot) \geq 0} \int_0^2 \left[Za_2 - t(Z) - Z \cdot k(a_2 - s(Z))\right] f(Z) dZ$$

where $s(\cdot)$ is the cumulative quantity that $S$ is willing to supply at price $p$, $t(Z) := \int_0^Z s(p) dp$ is defined as the payment transferred from $S$ to $B$, i.e. $S$’s adverse-selection cost, and $f(z) = 1/2$ is the p.d.f. of $Z$. Write the Lagrangian as $L(Z, s, t) := (Za_2 - t - Zk(a_2 - s)) f(Z)$. (It is easy to verify that the Lagrangian is concave in $(s, t)$.) The Euler-Lagrange equation is

$$\frac{d^2}{dz^2} [-Zk_a(a_2 - s(Z))] f(Z) = -f(Z),$$

which is an ordinary differential equation, subject to the terminal conditions $s(2) = a_2$ ($S$ sells everything at the highest possible price $Z = 2$) and $t(0) = 0$. Off corner, the solution is simply $s(Z) = a_2 - k_a^{-1}(2Z - 1)$. With this solution, the Euler-Lagrange equation can be rearranged as $Zk_a(a_2 - s(Z)) = (1 - F(Z)) / f(z)$, the left-hand side of which is the marginal private value and the right-hand side of which is the inverse hazard rate. The feasibility constraint requires $s(Z) > 0$, which results in a corner solution that for $Z \leq \tilde{z}(a_2) := 2 / (k_a(a_2) + 1)$, $s(Z) = 0$. Finally, it is easy to verify that
Note that the effect of \( \frac{\partial}{\partial \beta} \) on \( \beta(m) \), which monotonically increases in \( m \). It therefore suffices to show \( \frac{\partial q_0}{\partial \beta} \) in equilibrium. The analysis in appendix 3.7.D shows that under condition (3.22), \( q_0 \) has interior solution, i.e. the first-order condition (3.21) holds as \( \alpha - (1 - \alpha - \beta) \bar{u}_\text{NL}'(a_0 - q_0) = 0 \). By implicit function theorem,

\[
\frac{\partial q_0}{\partial \beta} = \frac{\partial^2 \bar{u}_\text{NL}' / (\partial q_0 \partial \beta)}{\partial^2 \bar{u}_\text{NL}' / \partial q_0^2} = -\frac{\partial p_0 / \partial \beta - \phi + \bar{u}_\text{NL}'(a_0 - q_0)}{(1 - \alpha - \beta) \bar{u}_\text{NL}''(a_0 - q_0)}.
\]

The numerator corresponds to the three effects discussed right after proposition 3.3. The first two terms add up to zero \( (\partial p_0 / \partial \beta = \phi) \) and the third term is positive. The denominator is negative by concavity of \( \bar{u}_\text{NL}(\cdot) \). Thus, together with the minus sign before the fraction, \( \frac{\partial q_0}{\partial \beta} > 0 \), i.e. the more \( M \) there are, the more \( S \) sells to them ex ante.

**Proposition 3.4**

*Proof.* From equation (3.18),

\[
\frac{\partial w}{\partial q_0} = -(1 - p_0) + (1 - \alpha - \beta) \tilde{k}_\text{NL}(a_0 - q_0) = -(1 - p_0) + (1 - \alpha - \beta) \left( 1 - \bar{u}_\text{NL}'(a_0 - q_0) - \tilde{i}_\text{NL}(a_0 - q_0) \right),
\]

where the second equality follows from the identity that \( a_2 = \bar{u}_\text{NL}(a_2) + \tilde{k}_\text{NL}(a_2) + \tilde{i}_\text{NL}(a_2) \), implying \( 1 = \bar{u}_\text{NL}'(a_2) + \tilde{k}_\text{NL}(a_2) + \tilde{i}_\text{NL}(a_2) \). At equilibrium, the first-order condition for \( S \)'s ex ante problem implies that \( \bar{u}_\text{NL}'(a_0 - q_0) = \alpha / (1 - \alpha - \beta) \); see equation (3.21) and the associated discussion. Therefore, evaluating the above partial derivative at equilibrium gives, after substituting \( p_0 \) by the expression in equation (3.17), \( \frac{\partial w}{\partial q_0} = -(1 - \phi) \beta - (1 - \alpha - \beta) \tilde{i}_\text{NL} \). Note from equation (3.13) that \( \tilde{i}_\text{NL}(a) = (k_a(a) / (1 + k_a(a)))^2 > 0 \). Hence, \( \partial w / \partial q_0 \) is negative in equilibrium.

**Proposition 3.5**

*Proof.* It is equivalent to show that when \( \phi \) is very small, there exists some \( m \) that welfare function \( w(m) \) is not increasing. The simplest such example can be found at \( m = 1 \) and \( \phi \) so small that condition (3.22) is binding.

First, it shall be shown that at \( m = 1 \), a binding condition (3.22) implies \( q_0 \) is virtually zero (infinitesimally small and yet positive). Denote the binding limit of \( \phi \) to be \( \phi := \)
2/(2 + k_u(a_0))/(2\theta - \theta^2)$, following condition (3.22). Then from the expression of the equilibrium $q_0$ (appendix 3.7.D):

$$\lim_{\phi \downarrow \phi} q_0(\cdot) \big|_{m=1} = \lim_{\phi \downarrow \phi} \left[ a_0 - k_a^{-1} \left( \frac{1 - \beta}{\alpha} - 2 \right) \right] \big|_{m=1} = 0,$$

where the last equality follows the expression of $\alpha$: $\lim_{\phi \downarrow \phi} \alpha = 1/(2 + k_u(a_0))$ and the expression of $\beta$: $\beta(m)\big|_{m=1} = 0$. Thus, when $\phi$ hits the lower bound of $2/(2\theta - \theta^2)/(2 + k_u(a_0))$, $q_0 \downarrow 0$ at $m = 1$.

Next, evaluate some of the components in equation (3.19). The effect on the adverse-selection:

$$\frac{\partial \tilde{k}}{\partial \bar{\beta}} = \beta \cdot \left( \frac{1 - \phi}{\alpha \phi^2} \right)^2 k_{aa} \left( k_{a}^{-1} \left( \beta \cdot \frac{1 - \phi}{\alpha \phi^2} \right) \right)^{-1},$$

which reduces to zero at $m = 1$ (as $\beta \big|_{m=1} = 0$). From the proof of proposition 3.4:

$$\frac{\partial w}{\partial q_0} = -(1 - p_0) + (1 - \alpha - \beta) \left( 1 - \tilde{u}_{\text{NL}}'(a_0 - q_0) - \tilde{r}_{\text{NL}}'(a_0 - q_0) \right)$$

$$= (p_0 - \alpha - \beta) - (1 - \alpha - \beta) \tilde{u}_{\text{NL}}'(a_0 - q_0) - (1 - \alpha - \beta) \tilde{r}_{\text{NL}}'(a_0 - q_0)$$

$$= -(1 - \phi) \beta + \alpha - (1 - \alpha - \beta) \tilde{u}_{\text{NL}}'(a_0 - q_0) - (1 - \alpha - \beta) \tilde{r}_{\text{NL}}'(a_0 - q_0),$$

which further simplifies to $\partial w/\partial q_0 = -(1 - \alpha) \tilde{r}_{\text{NL}}'(a_0)$ as evaluated at $m = 1$ with $q_0 = 0$ and $\bar{\beta} = 0$.

With the above results, the partial derivative of welfare with respect to $\beta$ becomes

$$\frac{\partial w}{\partial \bar{\beta}} \big|_{m=1, \phi \downarrow \phi} = \tilde{k}_{\text{NL}}(a_0) - (1 - \alpha) \tilde{r}_{\text{NL}}'(a_0) \frac{\partial q_0}{\partial \bar{\beta}} \big|_{m=1, \phi \downarrow \phi}.$$

From the first-order condition (3.21), it can be computed that

$$\frac{\partial q_0}{\partial \bar{\beta}} \big|_{m=1, \phi \downarrow \phi} = -\frac{1}{1 - \alpha \tilde{u}_{\text{NL}}'(a_0)} (> 0).$$

Hence,

$$\frac{\partial w}{\partial \bar{\beta}} \big|_{m=1, \phi \downarrow \phi} = \tilde{k}_{\text{NL}}(a_0) + \tilde{r}_{\text{NL}}'(a_0) \tilde{u}_{\text{NL}}'(a_0) \tilde{u}_{\text{NL}}(a_0) = a_0 - \tilde{r}_{\text{NL}}(a_0) - \tilde{u}_{\text{NL}}(a_0) + \tilde{r}_{\text{NL}}(a_0) \tilde{u}_{\text{NL}}'(a_0) \tilde{u}_{\text{NL}}(a_0)$$

where the last equality follows the accounting identity that $\tilde{k}_{\text{NL}}(a) = a - \tilde{r}_{\text{NL}}(a) - \tilde{u}_{\text{NL}}(a).$ Note that $-\tilde{r}_{\text{NL}}(a_0) < 0$ and that $\tilde{u}_{\text{NL}}'(\cdot)$ is concavely increasing, hence implying $-\tilde{u}_{\text{NL}}(a_0) <$
$-\hat{a}_{NL}(a_0)a_0$. Plugging in the expressions for $\hat{t}'_{NL}(a_0)$, $\hat{u}'_{NL}(a_0)$, and $\hat{u}''_{NL}(a_0)$ (equations 3.15 and 3.16), one can get

$$\frac{\partial w}{\partial \beta} < a_0 \cdot (1 - \hat{u}'_{NL}(a_0)) + \hat{u}'_{NL}(a_0) \frac{\hat{t}'_{NL}(a_0)}{\hat{u}'_{NL}(a_0)} - k_a(a_0) < 0$$

where the last inequality follows by the assumption that $k_{aa}(a_0) > 0$, which implies $k_a(a)$ is concavely increasing and hence $k_{aa}(a)a < k_a(a)$ (also note that $k_a(0) = 0$).

Therefore, when $\phi$ is small and hits the constraint implied by condition (3.22), $\partial w/\partial \beta < 0$ at $m = 1$. That is, a sufficiently large RO size $\phi$ is a necessary condition for a monotone increasing welfare function. \hfill $\square$

### 3.7.C Probability of an MM trade conditional on an interim trading price

This appendix evaluates the conditional probability of an MM trade, given an interim trading price $P_1 = p_1 \leq 2\phi$. Let $X$ be the indicator random variable which is 1 if the interim trade is MM and 0 otherwise. Then the interim trading price $P_1$ can be written as $P_1 = X\phi Z + (1 - X)Z$. The c.d.f. of $P_1$ is

$$P(P_1 \leq p_1) = \begin{cases} \gamma p_1/(2\phi) + (1 - \gamma)p_1/2, & \text{if } 0 \leq p_1 \leq 2\phi \\ \gamma + (1 - \gamma)p_1/2, & \text{if } 2\phi < p_1 \leq 2 \end{cases}$$

where $\gamma := P(X = 1) = \beta/(\beta + \alpha)$. Suppose the realization is $p_1 \leq 2\phi$, the conditional MM trade probability is

$$\hat{\beta} = P(X = 1|P_1 = p_1, P_1 \leq 2\phi) = \frac{P(X = 1, P_1 = p_1|P_1 \leq 2\phi)P(P_1 \leq 2\phi)}{P(P_1 = p_1|P_1 \leq 2\phi)P(P_1 \leq 2\phi)}.$$

The denominator is the conditional (on $P_1 \leq 2\phi$) density of $P_1$, which is $1/(2\phi)$ as easily derived from the c.d.f. above:

$$P(P_1 = p_1|P_1 \leq 2\phi) = \frac{\partial}{\partial p_1} \frac{P(P_1 \leq p_1, P_1 \leq 2\phi)}{P(P_1 \leq 2\phi)} = \frac{\partial}{\partial p_1} \frac{\gamma p_1/(2\phi) + (1 - \gamma)p_1/2}{\gamma + (1 - \gamma)p_1/2} = \frac{1}{2\phi}.$$

The numerator can be further computed by law of total probabilities:

$$P(X = 1, P_1 = p_1|P_1 \leq 2\phi) = P(P_1 = p_1|X = 1, P_1 \leq 2\phi) \cdot P(X = 1|P_1 \leq 2\phi).$$

The first part $P(P_1 = p_1|X = 1, P_1 \leq 2\phi)$ is simply the density of a random variable uniformly distributed on $[0,2\phi]$, i.e. $1/(2\phi)$. The second part is $P(X = 1|P_1 \leq 2\phi) = P(P_1 \leq 2\phi)$.
2\phi|X = 1)\mathbb{P}(X = 1)/\mathbb{P}(P_1 \leq 2\phi) = 1 \cdot \gamma / (\gamma + (1 - \gamma)\phi). \) Assembling all these parts, one can get \( \hat{\beta} = \beta / (\beta + \alpha\phi) \) as in equation (3.5). Note that the expression does not depend on the specific realization of \( P_1 \), due to the uniform distribution of \( Z \). For a generic specification of \( Z \), this conditional probability will be a function of \( p_1 \) and can complicate the analysis substantially.

### 3.7.D Solution to \( S \)'s ex ante optimization problem

After simplifying, the optimization problem can be written as

\[
\hat{u}_0^S = \max_{q_0} p_0 q_0 + \alpha \cdot \left[ a_2 - \phi^2 k(a_2) \right]
\]

(3.20)

\[
\hat{u}_0^S = \beta \cdot \left[ a_2 + (1 - \phi) k_2^{-1}(a_2) \right] + (1 - \alpha - \beta) \hat{u}_{NL}^S(a_2),
\]

where \( a_2 = a_0 - q_0 \). There are two possible interior solutions for \( q_0 \), depending on whether \( k_a(a_0 - q_0) \leq \tilde{k}_a \), as suggested in the first-order condition:

\[
\alpha + \left[ \alpha \phi^2 k(a_0 - q_0) - \beta \cdot (1 - \phi) \right] \mathbb{I}_{k_a(a_0 - q_0) < \tilde{k}_a} - (1 - \alpha - \beta) \hat{u}_{NL}^S(a_0 - q_0) = 0,
\]

where \( \hat{u}_{NL}^S(a) = \mathbb{E}u_{NL}^S(a, Z) \) as given in equation (3.11). In particular, \( \hat{u}_0^S(a) = 1 / (1 + k_a(a)) > 0 \) and \( \hat{u}_{NL}^S(a) = -k_{av}(a)/(1 + k_a(a))^2 < 0 \). These guarantee that the second-order condition is satisfied, i.e. \( \hat{u}_0^S(q_0) \) is concave in \( q_0 \).

There are, however, three corners to the first-order condition. First, consider the lower bound of \( q_0 \geq 0 \). If this corner binds, i.e. if the left-hand side of the first-order condition (3.21) is negative at \( q_0 = 0 \), then by the concavity of \( \hat{u}_0^S(a) \), \( S \) never wants to sell anything in the ex ante period. To rule out this uninteresting case, assume the opposite, i.e. \( \hat{u}_0^S(q_0)/dq_0|_{q_0=0} > 0 \). Suppose, for now, that \( k_a(a_0) \geq \hat{k}_a \) (verified below) and note that \( \beta \geq 0 \). Then a sufficient constraint, from the first-order condition (3.21), is \( \hat{u}_0^S(q_0)/dq_0|_{q_0=0} = \alpha - (1 - \alpha - \beta) \hat{u}_{NL}^S(a_0) \geq \alpha - (1 - \alpha) \hat{u}_{NL}^S(a_0) > 0 \), or after simplification,

\[
(2\theta - \theta^2) \phi > 2 / (2 + k_a(a_0)),
\]

(3.22)

where \( 0 < \phi < 1 \) and \( 0 < \theta < 1 \). To verify that constraint (3.22) is a sufficient condition for \( k_a(a_0) > \hat{k}_a \), first note from condition (3.22) that \( (2\theta - \theta^2) \phi > 2/3, 2\theta - \theta^2 > 2/3, \) and \( \phi > 2/3 \) because \( 0 < k_a(a_0) < 1 \) and \( 0 < 2\theta - \theta^2 < 1 \). Second, from the expression of \( \beta \) it follows that \( \beta < (1 - \theta)^2 = 1 - (2\theta - \theta^2) < 1 - 2/3 = 1/3; \) and hence, \( \phi^2 - \beta > (2/3)^2 - 1/3 = 1/9 > 0 \). Then rewrite (3.22) as \( k_a(a_0) > 1/\alpha - 2 \) and subtract both sides \( \hat{k}_a \) to get

\[
k_a(a_0) - \hat{k}_a > 1 / \alpha - 2 - \hat{k}_a = \frac{2}{2\theta - \theta^2} \frac{1}{\phi^3} (\phi^2 - (1 - \phi)\beta) - 2 > 2 \cdot \frac{1 - \phi}{\phi^3} (\phi^2 - \beta) > 0
\]
where the first equality follows by substituting the expression of $\alpha = (\theta - \theta^2/2)\phi$ and the last but second inequality follows because $2/(2\theta - \theta^2) > 2$ for all $0 < \theta < 1$ (note that $\phi^2 - \beta > 0$ implies $\phi^2 - (1 - \phi)\beta > 0$).

Next, consider the in-between bound of $k_a(a_0 - q_0) \leq \hat{k}_a$. Note that condition (3.22) implies $k_a(a_0) > \hat{k}_a$. Then there is some $q > 0$ such that $a_0 - q = k_a^{-1}(\hat{k}_a)$ by the monotonicity of $k_a(a)$. Check $\partial \tilde{u}_S^S/\partial q_0$ at such a $q$, $\alpha - (1 - \alpha - \beta)u_{NL}^S(k_a^{-1}(\hat{k}_a))$, which is strictly increasing in $m$ because both $\beta(m)$ and $\hat{k}_a$ are increasing in $m$. Therefore, it suffices to check whether in the limit of $m \to \infty$ this derivative is positive or negative: If the limit is negative, by concavity of $\tilde{u}_S^S$, the unique optimal $q_0$ is strictly smaller than $k_a^{-1}(\hat{k}_a)$ and this in-between bound never binds. Let $\beta \to (1 - \theta)^2$ and the above derivative can be simplified to get that it is negative if and only if $(2\theta - \theta^2)(1 + \phi^2) > 1$. Then $(2\theta - \theta^2)(1 + \phi^2) = (2\theta - \theta^2) + ((2\theta - \theta^2)\phi)\phi > \frac{2}{3} + \frac{2}{3} \cdot \frac{2}{3} = \frac{10}{9} > 1$. That is, condition (3.22) also guarantees that in equilibrium, $k_a(a_0 - q_0) > \hat{k}_a$.

Finally, consider the upper bound of $q_0 \leq a_0$. S wants to sell all her position only if $\partial \tilde{u}_S^S/\partial q_0$ evaluated at $q_0 = a_0$ is non-negative: $\alpha - (1 - \alpha - \beta) = 2\alpha + \beta - 1 \geq 0$. This inequality, however, never holds because $2\alpha = (2\theta - \theta^2)\phi < 1 - (1 - \theta)^2$ and $\beta < (1 - \theta)^2$. Therefore, the upper bound is never binding: S never sells everything to M in the ex ante period.

To conclude, with condition (3.22), the ex ante equilibrium supply by S has a unique interior solution $q_0 = a_0 - k_a^{-1}((1 - \beta)/\alpha - 2)$. 


Opening the black box of how orders are submitted, queued, and processed

Queuing Uncertainty

Queuing of limit orders is the focus of this chapter: In a high-speed trading environment, traders simultaneously react to public information not knowing the sequence in which their orders arrive at the exchange. A theoretical model is developed to capture such queuing uncertainty. Market makers strategically compete for rents in liquidity provision. In equilibrium, liquidity first overshoots—orders at the end of the queue make expected losses—and then is immediately withdrawn, resulting in “flickering orders”. A boost in the trading speed amplifies the overshoot but the effect on order book dynamics (strategic order submission and cancellation) depends on the source of the speed increase. These predictions echo empirical evidence on and policy concerns over “quote stuffing”, order-to-trade ratios, and minimum quote life. The model points to an optimal level of queuing uncertainty, to which the exchange can steer by carefully randomizing the limit order queues.
This chapter is based on Yueshen (2014). It benefits tremendously from very helpful discussions with Mark van Achter, Shmuel Baruch, Eric Budish, Jérôme Dugast, Terrance Hendershott, Bengt Holmström, Johan Hombert, Bernard Hosman, Charles Jones, Vincent van Kervel, Andrei Kirilenko, Andrey Malenko, Katya Malinova, Albert Menkveld, Sophie Moinas, Andreas Park, Barbara Rindi, Ioanid Roșu, Spyros Skouras, Kumar Venkataraman, Adrien Verdelhan, Mao Ye, Haoxiang Zhu, and, in particular, the seminar and conference participants at the University of Gothenburg, Rotterdam School of Management, Aalto University Business School, INSEAD, IHS Vienna, Paris-Dauphine, 11th International Paris Finance Meeting, SAEe 2013 (Santander), MBF Conference (Rome), and NBER Market Microstructure Meeting 2013.

4.1 Introduction

Since the automation of securities trading in the late 1990s, speed has become a salient feature of modern financial markets. The rise of the machines—technological advancement marked by the upsurge of algorithmic and high-frequency trading—has remarkably reduced trading latency from minutes (Biais, Hillion, and Spatt, 1995) to milliseconds (Hasbrouck and Saar, 2013), with the latest record at nanoseconds (Gai, Yao, and Ye, 2013).

As the trading speed approaches the limit imposed by nature, it is arguable whether market participants can still perfectly condition their decisions on the real-time market status. For example, when a trader sees an update of the limit order book, is this the true book status at that moment? Probably not: Among many potential delays, it takes time for the electronic message to travel from the exchange to the trader, and, during this period, the order book might evolve.

Under the premise that traders cannot perfectly condition on the true market status in real-time (especially in a high speed trading environment), this paper formally addresses the following questions: What does such imperfect conditioning imply for market participants’ optimal decisions? What are the implications for liquidity provision? How to evaluate market quality under such imperfection? How should the market be organized and what should regulators do, if anything, to deal with the consequences?

Threefold contributions are put forward in this paper, with a focus on the liquidity provision by limit order traders in an electronic limit order book. First, a novel theoretical framework is developed to analyze optimal limit order submission, accounting explicitly for traders’ imperfect conditioning on the market status. Second, the model yields rich testable predictions, echoing evidence from the existing literature, on equilibrium liquidity provision and dynamics, e.g. “flickering depth”, “quote stuffing”, heavy electronic message

\footnote{A photon travels in a vacuum from New York City to Chicago in about four milliseconds, an eternity in a trading environment with latency at nanoseconds.}
Figure 4.1: Game structure comparison. This figure illustrates the difference between the conventional view and the new view (this paper) for low-latency or high-speed trading. In panel (a), agents arrive sequentially and react to what previous agents have done. Panel (b) adds queuing uncertainty to panel (a). Nature queues the two agents randomly, but they do not know their queue positions. The agents in panel (b) effectively play a simultaneous game.

traffic (in absence of fundamental shocks), and “holes” in limit order books. A competition channel is identified through which different types of trading latencies affect strategic liquidity provision differently. Finally, market quality is examined through a welfare criterion embedded in the model. The generated policy implications anchor recent regulatory debates regarding high-frequency trading, e.g. minimum order resting time, capping order-to-trade ratio, batching orders and randomizing queues, etc. It is shown that trading efficiency can be improved by adjusting—sometimes increasing—the uncertainty in the market.

The paper features a novel game structure to model trading games. Figure 4.1 illustrates the idea. Consider a very short time period immediate after an information event. Panel (a) presents the extensive form of the classical treatment of such a scenario, where the sequential nature of the game guarantees that each agent can perfectly condition on the latest market status by observing what previous agents have done (a perfect information sequential game). Panel (b) contrasts this classical view with “queuing uncertainty”, a specific form of the market imperfection discussed above. Nature queues the agents randomly, and the agents do not observe their queue positions (note the information sets shown by the dashed lines). Hence, the true market status cannot be perfectly conditioned on.\footnote{Model-wise, the game described in panel (b) is simultaneous, though it should be emphasized that the}
The model builds on the seminal work of Glosten (1994), adding explicitly the limit order submission process with the above queuing uncertainty. Glosten (1994) characterizes an equilibrium limit order book by invoking a zero-profit condition at each price level (with infinitely many limit order traders; zero rent, perfect competition). Relatedly, Sandås (2001) constructs and tests an equilibrium with a set of break-even conditions, i.e. the marginal (last) limit order earns zero expected profit at each price level. Queuing uncertainty contrasts with the two equilibrium conditions: While ex ante of queue realization, traders expect to split a positive rent, ex post, the marginal (last) limit order at each price level always makes losses (not breaking-even).

Aggregate liquidity always “overshoots” under such equilibrium with queuing uncertainty: Not all infra-marginal limit orders earn positive expected profit; those at the end of the queue, regardless of the queue realization, always lose. The reason is that the traders optimally choose their limit order sizes such that the marginal orders break even individually in expectation, where this expectation is taken over all possible queue positions. With queuing uncertainty, an order in the top of the queue earns positive profit but a bottom order loses (as execution probability lowers and adverse-selection cost increases along the queue\(^1\)) and for any queue realizations, there always is one (unfortunate) bottom limit order. Effectively, the limit order traders engage in strategic competition, as each trader’s order size negatively affects the expected profitability of all others’. Queuing uncertainty weakens the strategic substitution effects and, consequently, amplifies the competition among the traders, leading to liquidity overshoot. To this extent, queuing uncertainty governs the level of competition.

Such liquidity overshoot is transitory and “unstable” As some limit order traders eventually realize that their orders are queued at the bottom and losing, the excess part will be canceled (if yet not executed). In a dynamic extension with repeated order modification, the model shows that such a revision option exacerbates the short-run overshoot: The traders compete more recklessly by submitting larger orders because they will be able later to cancel (revise) the orders that “unfortunately” sit in the bottom of the queue.

The dynamics of the displayed order book is described by the limit order traders’ equilibrium order submission and revision: There is intense overshoot of book depth in the short-run and then it gradually dies out in the long-run. The process complements Glosten (1994) and Sandås (2001) by showing how a “stable” level of book depth is achieved over time. The model further reveals that drops in trading latency (faster trading speed) amplify the liquidity overshoot in the short-run due to better revision options, and it takes many more agents are not required to move simultaneously in the real world. Simultaneity is a modeling technique to reflect queuing uncertainty.

\(^1\) The decreasing marginal profitability along the queue inherits from Glosten (1994). See also Hollifield, Miller, and Sandås (2004), Hasbrouck and Saar (2009), Liu (2009), and Raman and Yadav (2013) for empirical evidence.
rounds of revisions for the book depth to revert to its “stable” level. To this extent, the model explains how trading latency drops can lead to an increased cancellation/execution ratio (e.g. Gai, Yao, and Ye, 2013) and argues that the so-called “quote stuffing” might be an equilibrium outcome resulting from traders’ inability to perfectly condition on real-time market conditions. This point is inspired by and similar to Baruch and Glosten (2013) who describe a mix-strategy equilibrium in a continuous price order book. The model framework can be readily used to formally analyze regulations concerning capping quote-to-trade ratios (implemented on, e.g., London Stock Exchange, Eurex Germany, Borsa Italiana among others) and minimum order resting time and the reasons why such policies may dampen liquidity provision.

However, latency drops can have ambiguous effects on the stabilization time of the order book. The model identifies two types of latencies. On the one hand, when agents’ reaction latency reduces, they are able to send more messages per unit of time, and hence, processing all orders takes longer time. On the other hand, when transmission latency reduces, the round-trip time of each order reduces, and the order book updates more frequently. All else being equal, the first type of latency reduction (for example, trading algorithm improvement and CPU upgrades) slows down the stabilization process, while the second (exchange server or optical fiber/microwave technology upgrades) hastens the convergence.

Although transitory, liquidity overshoot is nevertheless very important because it typically occurs immediately after an information event (a news announcement or simply an update in the order book) and is followed by market orders trading on such information. Liquidity demanders (fundamental investors) benefit from the overshoot as they can meet their needs at lower cost (c.f. Glosten, 1998). Such welfare concern echoes recent debate on how to regulate trading speed. Larry Harris suggests that “[r]egulatory authorities could require that all exchanges delay the processing of every posting, canceling and taking instruction they receive by a random period of between 0 and 10 milliseconds.” 

Correspondingly, EBS, a leading foreign exchange trading platform, recently implemented queue randomization: “[M]essages transmitting orders...will be bundled into batches and then run through a process that randomizes their place in the queue” and the randomization takes “between one and three milliseconds”.2

As a model application, this paper formally examines the above proposal, especially how queue randomization affects market quality. Intuitively, welfare generally improves with market liquidity, suggesting that adding queuing uncertainty to the market might improve efficiency because elevated competition will amplify the liquidity overshoot. However, there are cases where welfare is inverse U-shaped in the aggregate liquidity supply: The reduction in the rent of market makers, due to excessive liquidity overshoot, might overcast the increase in liquidity demander’s gain. By properly randomizing the queues—essentially

---

adjusting the degree of competition among liquidity providers—the exchange can steer the aggregate liquidity supply toward the socially optimal level. Importantly, such randomization should be carried out with caution because 1) the effectiveness of randomizing queues is bounded by the ex ante speed heterogeneity of all traders (and can be very limited); 2) the optimal randomness in queuing is a delicate measure and can be hard to determine in real-time trading; and 3) by adding queuing uncertainty, the exchange essentially deprives fast traders of their market power, disincentivizing participation in the long-run.

Additional related literature is discussed in section 4.2. A baseline model is developed in section 4.3. Section 4.4 explores the time dimension extensions and analyzes the book depth dynamics. Section 4.5 generates implications on allocative efficiency and market design. On the price dimension, an analysis of liquidity overshoot deep in the limit order book is provided in section 4.6. Section 4.7 then concludes. Appendix 4.8.A provides a tractable micro-foundation for queuing uncertainty based on speed heterogeneity. Appendix 4.8.B summarizes the notations. All proofs are collated in appendix 4.8.C.

4.2 Related literature


The issue of uncertain sequencing in limit order market previously focuses on the market orders and their execution. For example, Foucault, Kozhan, and Tham (2013) study “toxic arbitrage” activities in a foreign exchange market setting where triangular arbitrage opportunities are exploited by fast traders, hurting dealers who have uncertainty revising their quotes timely. In a related setting, Budish, Cramton, and Shim (2013) argue that the continuous time market design of financial asset trading is flawed in that the market makers’ inability to synchronize stale quotes with innovations in a reference asset results in a prisoners’ dilemma of technology arm race. van Kervel (2013) studies liquidity provision by high-frequency traders who duplicate limit orders across venues and cancel immediately after trades occur in one of the venues. Hence, market order traders may or may not consume the visible liquidity. To compare, the current paper features the queue of limit orders, while the market order arrival is modeled to close the model.

Queuing uncertainty is also similar to the “contact-order uncertainty” in opaque over-the-counter (OTC) markets, as modeled in Zhu (2012). Despite the difference in the mar-
4.2. RELATED LITERATURE

In the market setting (limit order book v.s. OTC market), in both models, liquidity providers do not know their queue positions. In the OTC setting, dealers compete strategically on their price quotes. In the limit order book setting, as the price grid is fixed in the book, the market makers compete on the size of their orders. The depth of the order book is, therefore, the focal liquidity measure of the current paper.

The model’s prediction on book depth dynamics recalls Baruch and Glosten (2013), who focus on the dynamics of quote prices in a limit order book with zero tick size, due to which liquidity providers can always undercut each other, resulting in a mixed-strategy equilibrium with “flickering quotes” (see Hasbrouck (2013) for an empirical analysis). Under the mixed-strategy equilibrium of their model, consistent with the current paper, agents’ decisions cannot be perfectly conditioned on by one another. Instead of quote prices, the current paper features short-run depth overshoot followed by immediate cancellation. Such “flickering depth” prediction agrees with Baruch and Glosten (2013) in that, the so-called “quote-stuffing” behavior by high-frequency traders, rather than gaming the trading system, might in fact be an equilibrium pattern as liquidity providers cancel their losing orders, the occurrence of which is nothing but an “unfortunate” realization of the orders’ queue position. In particular, this paper shows that, given sufficient time for revision, the limit order book converges to the stable equilibrium described in Sandås (2001) (in contrast to Back and Baruch (2013), such convergence does not require infinite market makers).

Liquidity overshoot is evidenced in Sandås (2001), who shows that the break-even condition does not hold on average and, instead, the marginal limit order loses. Such losses are attributed to the negative order processing cost (see tables 5 and 6 in Sandås, 2001), which is interpreted as a result of traders’ heterogeneous valuations for the asset. Queuing uncertainty provides an alternative equilibrium explanation that does not require heterogeneity in traders.

The current paper also adds to the debate on the impact of trading speed, e.g. due to the high-frequency traders, on market quality. See, among others, Pagnotta and Philippou (2013), Biais, Foucault, and Moinas (2013), Hoffmann (2014), and Menkveld and Zoican (2013) for related studies. Notably, the current model examines market quality only in terms of the gains from trade (competition on the intensive margin) and does not rely on the technology arm race (competition on the extensive margin), which is potentially socially costly. Accompanying the theory literature, empirical evidence on how low-latency trading technology affects market quality is also growing. See Hendershott, Jones, and Menkveld (2011) on the effect of the autoquote system on NYSE market quality; Jovanovic

---

1 The drivers, however, are different: It is the zero tick size assumption in Baruch and Glosten (2013), while in the current paper it is the queuing uncertainty resulting from the non-zero latency at which any trader can register new market updates. The current paper, hence, does not have direct implication on quote prices, while Baruch and Glosten (2013) do not generate predictions on order book depth (rigorously speaking, depth is not properly defined in a zero tick size limit order market).
and Menkveld (2012) on the competition between an entrant platform, Chi-X, and the incumbent Euronext in Belgian and Dutch equity markets; Riordan and Storkenmaier (2012) on the speed improvement of Deutsche Boerse in 2007; Chaboud et al. (2013) on the impact of high-frequency trading on foreign exchange markets; and Hasbrouck and Saar (2013) on low-latency trading strategy in NASDAQ.

The model also emphasizes the important difference of the effects of latency drops on market quality. For example, while Hasbrouck and Saar (2013) find order book tends to be deeper when reaction latency is low (large amount of low-latency activity), studies on exchange speed upgrades—Riordan and Storkenmaier (2012) and Gai, Yao, and Ye (2013), for example—find lower transmission latency is associated with reduced displayed book depth. On colocation, Brogaard et al. (2013) and Frino, Mollica, and Webb (2013) find that upgrades and introduction both deepen the order book. The distinction in the types of latency drops is unclear in the above literature, as, for example, traders’ reaction latency probably correlate with the exchange’s transmission latency. Empirical methods to disentangle the effects are expected.

4.3 Baseline model

This section presents a baseline model to illustrate the driver of the results of this study. The focus is on the liquidity provision at a given price level in a very short period, immediately following some public information event. Such information event could be macro news, earnings announcement, or simply an update in the order book. More concretely, for example, when a sell market order hits the best bid price, the updated asset value typically drops and market makers might want to add ask orders at the best ask price. The sell market order is the information event and the best ask price is the battlefield analyzed in this section. The one-price focus is motivated by the empirical fact that there are few market orders that bite deep into the book, and most of the order activity concentrates on the top of the limit order book. The dynamic extension is studied in section 4.4 and the extension to multiple price levels is considered in section 4.6.

4.3.1 Model setup

The model is consistent with Glosten (1994). The new element is queuing uncertainty, as explained below. This baseline model studies only the liquidity supply side, leaving the demand side exogenous. Sections 4.5 endogenizes liquidity demanders.

---

1 The equilibrium in Glosten (1994) assumes infinitely many market makers. Focusing on the strategic interaction, this section looks at finitely many market makers instead.
 Agents. There are \( n \) risk-neutral agents indexed by \( i \in \mathcal{N} := \{1, \ldots, n\} \), where \( n \) is a positive integer (the special case of a continuum of agents is considered in section 4.4). These agents represent the market’s liquidity supply side, which is the focus of this section, and they will be referred to as market makers (to be consistent with the literature), though in reality these agents can be any traders who want to use limit orders at such occasions (e.g. endogenous liquidity providers, or ELPs, as studied in Anand and Venkataraman, 2013).

Timing. The baseline model is a one-period simultaneous game. Each market maker “simultaneously” and independently chooses his limit order size, \( q_i (\geq 0) \). These orders are then queued by nature according to a known distribution (see “queuing uncertainty” below; appendix 4.8.A provides a tractable example for this distribution). Finally, payoffs realize (due to unmodeled exogenous arrival of market orders that trade against the market makers’ limit orders). The extensive form of the game with \( n = 2 \) market makers is illustrated in panel (b) of figure 4.1.

Limit order profitability. The limit orders are perfectly divisible. Let \( \pi(y) \) be the marginal profit (conditional on execution) of the \( y \)-th limit order at the given price level.\(^1\) Then the (expected) profit of the first limit order of size \( y \geq 0 \) at the given price is \( \pi(y) = \int_0^y \pi(x)dx \). For example, if a limit order of size 10 is added to a price level with existing depth of 5 units, it earns \( \int_5^{5+10} \pi(y)dy = \pi(15) - \pi(5) \) in expectation.

Assumption 4.1 (Quasi-top-of-queue advantage). The (expected) profit function \( \pi(y) \) is twice-differentiable and is quasi-concave on \( y \in [0, \infty) \). Define its first order derivative to be \( \dot{\pi}(y) \). The two terminal conditions hold: 1) \( \dot{\pi}(0) > 0 \) and 2) \( \lim_{y \to \infty} \dot{\pi}(y) < 0 \). Finally, normalize \( \pi(0) = 0 \).

Such a \( \pi(y) \) is assumed to be exogenously given in this section and the next. In section 4.5, \( \pi(y) \) will be endogenized and it will be shown that assumption 4.1 holds exactly. Some remarks are in order.

Remark 1. Quasi-concavity, together with the terminal conditions, imply that the marginal profit, \( \dot{\pi}(y) \), is first positive and then negative (but never positive again) as \( y \) increases. That is, if one thinks of a limit order book as a set of queues of orders (at various prices), orders on top of each queue earn profit while those in the bottom lose, hence the term “(quasi-)top-of-queue advantage”. If \( \pi(y) \) is strengthened to be concave, the top-of-queue advantage is exact: The marginal profitability always decreases along the queue.

Remark 2. A quasi-concave \( \pi(y) \) is reasonable because orders in the bottom of the queue have lower execution probability and are subject to higher adverse-selection costs (large

\(^1\)This paper adopts Newton’s notation, variables with dot(s) overhead, to indicate derivatives.
CHAPTER 4. QUEUING UNCERTAINTY

market orders tend to carry strong signals). See the literature listed in footnote 1 for empirical evidence for decreasing marginal profitability.

Remark 3. The two terminal conditions are assumed to rule out triviality. First, \( \dot{\pi}(0) > 0 \) guarantees participation: Otherwise, by quasi-concavity, \( \pi(y) \leq 0 \) for all \( y \geq 0 \) and each limit order at the price always loses. Second, \( \dot{\pi}(\infty) < 0 \) rules out the uninteresting case where the marginal profit is always strictly positive (inducing all market makers submit infinitely large limit orders).

Queuing uncertainty. Let \( \mathcal{K} \) be the collection of all possible queues.\(^1\) Ties (for example, both market makers 1 and 2 are at the same position in queue) are ruled out. Each queue is represented by an \( n \)-by-1 vector \( k \in \mathcal{N}^n \), where the \( i \)-th value \( k_i \in \mathcal{N} \) is the queue position of market maker \( i \). For example, if \( n = 3 \), a queue \( k = [2, 3, 1]^\top \) means market maker 1 is the second in the queue, market maker 2 the third, and market maker 3 the first. Fix a probability measure that defines queue distribution \( \mathbb{P}(K = k) \) for all \( k \in \mathcal{K} \). The distribution of the random vector \( K \) is known to all market makers.\(^2\)

4.3.2 Equilibrium

Consider market maker \( i \). He chooses his order size \( q_i \), given all others’ order sizes, to maximize his expected profit. He cares about the queue position of his order because of the top-of-queue advantage (assumption 4.1). For a given queue \( k \), write the aggregate size of the orders that queue before market maker \( i \)'s order by

\[
Q_i^<(k) = \sum_{j \in \mathcal{N}} q_j \mathbb{1}_{\{k_j < k_i\}},
\]

where the superscript “<” emphasizes that \( Q_i^< \) only counts the limit orders that queue strictly before \( i \)'s own order. A (pure-strategy) Nash equilibrium is a set \( q = \{q_1, \ldots, q_n\} \) such that \( \forall i \in \mathcal{N} \), given \( q_{-i} = q \setminus \{q_i\} \),

\[
q_i \in \arg \max_{q_i} \mathbb{E} \left[ \pi(Q_i^<(K) + q_i) - \pi(Q_i^< (K)) \right].
\]

The existence of such a Nash equilibrium is established by the following lemma.

Lemma 4.1 (Existence of Nash equilibrium). There exists a (pure-strategy) Nash equilibrium \( q \in [0, \tilde{y}]^n \) where \( \tilde{y} > 0 \) is the unique solution to \( \dot{\pi}(y) = 0 \) and at least one \( q_i \) is strictly positive.

\(^1\) \( \mathcal{K} \) is a finite set because there are at most \( n! \) possible queues (for \( n < \infty \)).

\(^2\) Note that the specification allows speed heterogeneity: Some market makers can be faster than others. For example, in a two-agent game, market-maker 1 is said to be faster than market-maker 2 if \( \mathbb{P}(K = [1, 2]^\top) > \mathbb{P}(K = [2, 1]^\top) \). Appendix 4.8.A characterizes such speed heterogeneity in terms of stochastic dominance.
4.3. BASELINE MODEL

Lemma 4.1 gives an upper bound, $\bar{y}$, for each individual’s equilibrium limit order size. The intuition is as follows: Note that $\bar{y}$ is the break-even point at which the marginal unit of limit order earns zero profit in expectation: $\dot{\pi}(\bar{y}) = 0$; see Sandås (2001) for a specific parametrization (e.g. equation 8). Therefore, no market maker will post a limit order exceeding $\bar{y}$ units because the part beyond the break-even point always loses (by quasi-concavity of $\pi(\cdot)$), regardless of the order’s queue position.

Clearly neither $q_i = 0$ for all $i \in \mathcal{N}$ or $q_i = \bar{y}$ for all $i \in \mathcal{N}$ ($n \geq 2$) is an equilibrium. At least one market maker’s order size is an interior solution, and the first-order condition holds for him:

$$\mathbb{E}\pi(Q_i^c(K) + q_i) = 0. \tag{4.2}$$

This first-order condition implies the following proposition.

**Proposition 4.1 (Liquidity overshoot).** In equilibrium, liquidity overshoots in the sense that the last unit of the limit order earns negative expected profit, $\pi(\Sigma_i q_i) \leq 0$. The equality holds if and only if there is one market maker who is almost surely the first in the queue, i.e. $\exists i \in \mathcal{N}$ such that $\mathbb{P}(K_i = 1) = 1$.

The intuition of this proposition is sketched in figure 4.2 with $n = 2$ market makers, whose orders are equally likely to be the first or the second in the queue. The downward sloping curve is the marginal profit $\dot{\pi}(\cdot)$. It crosses the horizontal axis at $\bar{y}$, which maximizes the profit $\pi(\cdot)$. If there were only one market maker (who almost surely arrives first in queue), this $\bar{y}$ would be the optimal quantity he would choose (c.f. the break-even condition of Sandås, 2001). The two market makers choose a (symmetric) equilibrium strategy $q^*$ such that each market maker has a zero expectation for his marginal profit, where the expectation is taken over all possible queue realizations. The first-order condition (4.2) simplifies to $\frac{1}{2}\dot{\pi}(q^*) + \frac{1}{2}\dot{\pi}(2q^*) = 0$. That is, they choose $q^*$ such that the absolute value of $\dot{\pi}(q^*)$ and $\dot{\pi}(2q^*)$ are the same. Clearly, by quasi-concavity of $\pi(\cdot)$, the marginal profit of the very last unit order in the book must be negative. In aggregate, therefore, liquidity overshoots.

The shaded area in figure 4.2 shows the size of the overshoot inefficiency. The key friction that creates such inefficiency is *queuing uncertainty*. If the queue is deterministic (panel (a) of figure 4.1), then the first market maker in the queue will submit an order of size $\bar{y}$ and the maximum aggregate profit is achieved. It should be noted that such inefficiency, however, is not necessarily socially costly: The overshoot of liquidity might benefit the demand side, which so far remains exogenous. Section 4.5 will endogenize the liquidity demand, and hence allow welfare analysis.

---

1 The break-even condition should not be confused with the related zero-profit condition. The former says the marginal limit order earns zero in expectation, while the latter says each market maker earns zero in expectation (perfectly competitive). Zero-profit for market makers is an asymptotic result as the number of market makers approaches infinity, as is the case in Glosten (1994) (see the proof of proposition 2 there).
CHAPTER 4. QUEUING UNCERTAINTY

Figure 4.2: Equilibrium illustration with $n = 2$ market makers. This graph illustrates the optimal supply decisions with two market makers, each having probability half of being the first in the queue. The downward sloping curve is the marginal profit function $\pi(q)$. It crosses $\pi(q) = 0$ at $q = \bar{y}$. In a symmetric-strategy equilibrium, the first-order condition (4.2) requires that the expected marginal profit be zero. That is, $\frac{1}{2} \pi(q^*) + \frac{1}{2} \pi(2q^*) = 0$, which implies that $\pi(q^*) = \pi(2q^*)$. The shaded area represents the (expected) loss that the last-in-queue market maker suffers.

To better understand the intuition of proposition 4.1, note that the game is reminiscent of oligopolistic competition among producers of a homogeneous good: The supply of one producer (market maker) negatively affects the profit of his competitors. Indeed, the liquidity provision in limit order market can be viewed as strategic substitutes, as defined in Bulow, Geanakoplos, and Klemperer (1985):

**Proposition 4.2 (Strategic substitution).** Suppose $\pi(\cdot)$ is concave on $[0, \sum_i q_i^*]$ where \{q_i^*\} is an equilibrium with $n \geq 2$ market makers. Then the market makers’ limit orders are strategic substitutes and, further, the substitution rate is lower than unity in absolute value. Mathematically, $\frac{\partial^2 E\pi(Q_i^r + q_i)}{\partial q_i \partial q_j} \leq 0$ and $\partial q_i^* / \partial q_j \in (-1, 0]$ for all $i, j \in \{1, 2, \ldots, n\}$ and $i \neq j$.

**Remark.** Note that quasi-concavity is strengthened to concavity to guarantee a monotone decreasing $\pi(\cdot)$ on the relevant support, akin to a monotone decreasing demand function.

Strategic substitutes means whenever a market maker $i$ increases his supply, he “threatens” to capture the others’ profit (when he queues to the front). However, the queuing uncertainty bounds the substitution rate. This is because other market makers “under-weigh” such threats as there is always non-zero probability for market maker $i$’s order to queue behind his competitor’s, in which case the “threat” is ineffective. As such threats are “underweighed”, the market makers engage in fierce competition which results in liquidity overshoot; that is, the last unit of the limit order no longer breaks even but loses. The key friction is queuing uncertainty, which leads to the intensified competition.
4.3.3 Examples

The baseline model above can be solved in closed-form under some specific parametrization. This subsection considers several of these examples.

**Quadratic profit function with symmetric market makers.** Consider the following: 1) Approximate the profit function by a second-order Taylor expansion and 2) assume that market makers are homogeneous—all queue realizations are equally likely and they use symmetric strategies. That is, assume that \( p(y) = \bar{y} - y \), for all \( y \geq 0 \) and that \( \mathbb{P}(K = k) = 1/(n!) \) for all \( k \in \mathcal{K} \).

The first-order condition (4.2), under symmetric strategy restriction, simplifies to
\[
\sum_{k=1}^{n} \frac{1}{n} \bar{\pi}(kq) = \frac{1}{n} \sum_{k=1}^{n} (\bar{y} - kq) = \bar{y} - \frac{n+1}{2} q = 0,
\]
and the unique symmetric, pure-strategy Nash equilibrium is that all market makers choose \( q^* = 2\bar{y}/(n+1) \), and the resulting total book depth is \( nq^* = 2n\bar{y}/(n+1) \). Note that though \( q^* \leq \bar{y} \), the total depth weakly exceeds \( \bar{y} \) (because \( 2n/(n+1) \leq 1 \)) and the equality holds only at \( n = 1 \).

Under this equilibrium, it can be computed that the strategic substitution effect is \( \partial q^*_i / \partial q_j = \mathbb{P}(K_j < K_i) = (n-1)/(2n) \in [0,1/2] \), always less than unity.

**Two market makers: one faster and one slower.** In reality some market makers (e.g. colocated high-frequency traders) might be faster than others. Let \( n = 2 \) and \( \mathbb{P}(K_1 = 1) = \alpha \), where \( \alpha \in [1/2, 1] \). Note that by construction, \( \mathbb{P}(K_2 = 1) = 1 - \alpha \in [0,1/2] \). In this case, market maker 1 is the faster one as his order is more likely to be queued in the front of market maker 2’s order. Then the first-order condition (4.2) becomes
\[
\begin{align*}
\alpha \pi(q_1) + (1 - \alpha) \bar{\pi}(q_2 + q_1) &= 0 \\
(1 - \alpha) \bar{\pi}(q_2) + (1 - \alpha) \pi(q_1 + q_2) &= 0
\end{align*}
\]
and with a second-order Taylor approximation, it can be solved that \( q^*_1(\alpha) = \alpha \bar{y} / (1 - \alpha + \alpha^2) \) and \( q^*_2(\alpha) = (1 - \alpha) \bar{y} / (1 - \alpha + \alpha^2) \). Observe that \( q^*_1(\alpha) \) is monotone increasing in \( \alpha \). That is, the faster is market maker 1, the larger his order will be.

4.4 Book depth dynamics

This section builds on the baseline model to analyze the book depth dynamics over time: The market makers are allowed to revise their existing (unexecuted) limit orders. As before, the focus remains on the book depth at a single price level in a short time interval immediately following some information event. A heuristic discussion is provided first to illustrate the idea.
**Order book dynamics: Overshoot followed by immediate cancellation.** Consider the following simplified scenario: Three market makers simultaneously submit limit orders to compete for a profitable \( \pi(y) \), satisfying assumption 4.1. Suppose the three orders are of the same size, \( q (> 0) \). For this example, let \( q < \bar{y} < 3q \), that is, there is liquidity overshoot. After the submission, a random order queue is formed at the exchange server and as the queue is processed, three book updates are *sequentially* disseminated to the market makers.

When the first update is observed, the *displayed* depth becomes \( q \), while the “true” depth is \( 3q \). Seeing the update, one of the three market makers is “happy”\(^1\): His order is queued in the top, earning positive expected profit \( \pi(q) > \pi(\bar{y}) = 0 \). However, the other two market makers become “worried” as they, knowing that they are not the first, realize that their orders are now more likely to be in the bottom of the queue; that is, their marginal orders lose in expectation. Therefore, these market makers with unresolved queuing uncertainty have an incentive to revise down their order sizes to some \( q' \in (0, q) \), which amounts to two additional revision orders submitted to the exchange. In equilibrium, there will still be liquidity overshoot after the revision: \( q + 2q' > \bar{y} \) (proved later in the formal analysis). The “happy” market maker has no incentive to submit new orders or modify his existing order (he would like to increase the order size, but doing so will, in most of the exchange market, have his order moved to the end of the queue, losing time priority).

When the second update is observed, the displayed depth is \( 2q \), while the true depth is \( q + 2q' \). A second market maker is “happy” when it turns out that his order is second-in-queue and after the previous revision, it is earning positive profit in expectation \( (q + q' < \bar{y} < q + 2q') \). The other market maker is disappointed because his order is in the bottom, part of it losing, and he wants to cancel the losing part. Therefore, another revision order is submitted: The last-in-queue market maker cancels the losing part of his order, \( q + 2q' - \bar{y} \).

As the second update of the book fully resolves the queuing uncertainty, no further revisions occur. In total, six orders are submitted in this example: Three new orders followed by three immediate revision/cancellation orders. The displayed book depth first increases and then decreases:

\[
0 \rightarrow q \rightarrow 2q \rightarrow 3q \rightarrow (2q + q') \rightarrow (q + 2q') \rightarrow \bar{y} \quad \text{(book stabilized)}
\]

book deepening \hspace{1cm} \text{book thinning}

Note that from \( 2q \) until \( \bar{y} \), there is liquidity overshoot.

\(^1\) In reality, inferring queue position can be tricky. After submission, limit order traders receive confirmation and book update separately from the exchange. The confirmation acknowledges with a time stamp that the exchange has processed the order. However, depending on the platform, the confirmation may (for example, in Eurex) or may not contain the order’s queue position. Even if there is no queue position in the confirmation, the trader can still (imperfectly) infer it by comparing the time stamps of the confirmation and the recent book updates. This section assumes perfect (yet sequential) inference of queue position. The case of no inference degenerates to the analysis in section 4.3, where there is no order revision after submission. These institutional insights are obtained from very helpful discussions with Andrei Kirilenko and with Bernard Hosman.
The rest of this subsection sets up a tractable framework that builds on this example to analyze the magnitude of liquidity overshoot, the number of revision orders, and the stabilization process of the order book. Apart from the results illustrated above, the model investigates how (different types of) trading latencies affect the equilibrium dynamics.

4.4.1 Model setup

The setup follows the baseline model. The top-of-queue advantage assumption 4.1 needs to be slightly strengthened:

**Assumption 4.2** (Top-of-queue advantage, strengthened). Assumption 4.1 holds and, further, \( \pi(y) \) is concave on \([0, \hat{y}]\) where \( \pi(\hat{y}) = 0 \).

This strengthened version of top-of-queue advantage is micro-founded by various commonly used frameworks in market microstructure literature, as shown in section 4.6 later. The next assumption simplifies the analysis by switching from an oligopolistic competition game to a perfect competition environment.

**Assumption 4.3.** There is a continuum of homogeneous market makers of mass 1.

**Remark 1.** The continuum of market makers has two implications. First, each market maker is of infinitesimally small market power (zero-measure mass) and therefore any individual’s strategy does not affect the aggregate depth. Second, because there are infinitely many market makers, the liquidity supply business is essentially in perfect competition: Market makers earn zero expected profit (shown below).

**Remark 2.** The homogeneity in market makers implies a specific distribution for each individual’s order queue position: \( K_i \) is i.i.d. uniform on \([0, 1]\) for all \( i \in [0, 1] \). It also allows the analysis to focus on symmetric equilibrium.

**Timing.** The time line is extended as below and is illustrated in figure 4.3. Time continuously runs from 0 to infinity. The game ends at time \( T \), the random arrival time of the market order. Suppose \( T \) has a c.d.f. of \( G(T) > 0 \) with a continuous support on \( T \geq 0 \).

**Two latencies.** The model distinguishes two different types of latencies. First, there is a gap \( \delta \ (> 0) \), referred to as the reaction latency, for each market maker to submit two consecutive orders (either addition or cancellation) to the exchange; that is, all market makers submit new orders only at time \( t \in \{0, \delta, 2\delta, \ldots \} \).

Second, there is transmission latency, \( \eta \ (> 0) \), defined below. Let the random round-trip time of each submitted order be i.i.d. from a uniform distribution with support \([0, \eta]\). Given the unit mass continuum of market makers (assumption 4.3), if a mass \( m \) of market makers
CHAPTER 4. QUEUING UNCERTAINTY

Figure 4.3: Time line of the dynamic game. This figure illustrates the dynamic game (section 4.4) played by the market makers. The market order can arrive at any time but the market makers operate at fixed latency, $\delta$. The probability that the market order arrives in between $t$ and $t + \delta$, conditioning on that it has not arrived by $t$, is denoted by $\phi(t)$. That is, $\phi(t) := (G(t + \delta) - G(t))/(1 - G(t))$. The queue is realized and revalued to the market makers incrementally every period.

all submit orders/revisions at time $t$, it takes (almost surely) $m/\eta$ units of time to observe all the book updates. That is, the amount of book updates observed is (almost surely) $dt/\eta$ per unit of time between $[t, t + m/\eta]$. To avoid complication of information asymmetry among market makers, each book update is assumed to arrive to all at the same time.$^1$

The reaction latency is motivated by the fact that it takes strictly positive amount of time—for example, CPU time—for each market maker to make a decision. Though admittedly the technology advancement has reduced the reaction latency to the level of nanoseconds, it has not reached zero (and neither will it). The transmission latency measures how soon, after an order is issued, the market participants can observe the update. It involves the delays of message travel time and the exchange server’s processing time. To sum up, market makers’ reaction latency determines the intensity at which orders are generated, while the transmission latency sets the rate at which the changes made by these orders are displayed.

To intuitively understand the difference between the two latencies, it may be useful to compare monitoring the order book to filming a movie. Figuratively, the reaction latency is like the frame rate (usually measured in frames per second, fps) of the movie: The lower is the latency, the more frames (order book changes) are generated each second. On the other hand, the transmission latency is like the “fast-forward” button that governs how fast the film is played, or how fast the order book changes are displayed.

**Time priority.** Time priority is enforced as before. That is, new orders submitted at later rounds will always be placed at the end of the existing queue. Modeled after the real-world practices, when the size of a limit order is revised down, its queue position is not affected; but when the size is revised up, the order moves to the end of the queue; for example, see *Conditions for Trading at Eurex Deutschland and Eurex Zürich*.

---

$^1$The setup here is consistent with Abreu and Brunnermeier (2003) who assume that in each instant of time there is a cohort of mass $1/\eta$ agents who becomes aware of mispricing in the economy. They refer to this period $[t, t + \eta]$ as the “awareness window”.
4.4. BOOK DEPTH DYNAMICS

4.4.2 Equilibrium

The focus is restricted to pure, symmetric-strategy equilibrium. Suppose the unit mass of market makers all submit an order of size \( q(0) \) at \( t = 0 \). Given the flow of book updates at rate \( 1/\eta \), the displayed order book depth is (almost surely) \( tq(0)/\eta \) for all \( t \in [0, \eta] \). Then, by \( t = h \delta, h \in \{0, 1, \ldots\} \) such that \( h \delta \leq \eta \), a mass of \( h \delta/\eta \) of the initial batch of orders’ queuing positions are revealed. Consequently, a mass of \( \min\{1, h \delta/\eta\} \) market makers observe the queue positions of their initial orders, while the other \( \max\{0, 1 - h \delta/\eta\} \) mass of market makers still face queuing uncertainty.

Denote the total rounds of order revisions by \( \bar{h} \); that is, for \( t \geq (\bar{h} + 1) \delta \), no market makers seek to revise their orders any more. (The value of \( \bar{h} \) will later be endogenously pinned down using the equilibrium conditions.) The following two results will be verified along the analysis below:

**Result.** 1) There is at least one revision, i.e. \( \bar{h} \geq 1 \). 2) For each \( h \in \{0, \ldots, \bar{h} - 1\} \), only the market makers whose initial orders are still subject to queuing uncertainty revise; the others do nothing.

The intuition behind these two results follows the liquidity overshoot (which persists throughout in equilibrium). For result 1), there are always some market makers who will find their orders end up in the bottom and make expected losses, and therefore, at least one round of cancellation is needed to revert the book depth to its stable level, \( \bar{y} \). Similarly, for result 2), as soon as a market maker has his order’s queuing uncertainty resolved, he no longer wants to revise it further: If his order queues before the break-even level of \( \bar{y} \), it earns positive expected profit and will stay; otherwise, the order is immediately fully canceled.

To this point, a competitive, symmetric, pure-strategy equilibrium can be concisely described by a path of \( \{q(h)\}_{h=0}^{\bar{h}-1} \), where each \( q(h) \) is the after-revision order size chosen by all active market makers, such that each market maker maximizes his per capita expected profit at each round \( h \). In particular, at \( h = 0 \), there is no revision and \( q(0) \) refers to the initial order size by all market makers. The terminal condition of \( q(\bar{h}) \) will be solved together with the value of \( \bar{h} \).

Fix a candidate equilibrium path of \( \{q(h)\} \). Recall that there is a mass of \( \delta/\eta \) market makers who resolve their queuing uncertainty each round. Then the true book depth—the depth computed by aggregating up all submitted orders—after the \( h \)-th revision becomes

\[
y(h) = \sum_{j=0}^{h-1} \int_{j \delta}^{(j+1) \delta} q_i(j) \text{d}i + \int_{h \delta \eta}^{1} q_i(h) \text{d}i = \sum_{j=0}^{h-1} \frac{\delta}{\eta} q(j)_{\text{stabilized}} + \left(1 - h \frac{\delta}{\eta}\right) q(h)_{\text{subject to revision}}.
\]

(The second equality holds by symmetry.) Note that the true depth is decomposed into a stabilized component and a to-be-revised component. The first component represents the
depth contributed by the inactive market makers whose queuing uncertainty has been resolved. The second component is the depth contributed by all active market makers who are still subject to queuing uncertainty and therefore, may revise in later rounds. The residual expected profit for the active market makers, given the stabilized component in the true depth (equation (4.3)), becomes

\[ \pi_h(y) = \pi\left(y + \sum_{i=0}^{h-1} \frac{\delta}{\eta} q(i)\right) - \pi\left(\sum_{i=0}^{h-1} \frac{\delta}{\eta} q(i)\right), \]

which is a left-down parallel shift of \( \pi(y) \). It inherits from \( \pi(y) \) a break-even point \( \bar{y}_h = \bar{y} - \frac{\delta}{\eta} \sum_{i=0}^{h-1} q(i) \) and satisfies assumption 4.2.

Consider next the optimization problem of the \((1-h\delta/\eta)\) mass of active market makers (whose initial orders are still subject to queuing uncertainty). To prepare for the Bellman equation, first construct an expression of their expected profit upon execution. By symmetry, each of these market makers knows that his own initial order will be queued at position \( K_h \), a uniformly distributed random variable on \([0, 1-h\delta/\eta]\). Upon execution, a market maker’s \textit{per capita} expected profit of his order, given all others’ order sizes, is \( \mathbb{E}\left[\pi_h(Q^<_i + q_i\Delta) - \pi_h(Q^<_i)\right]/\Delta, \) where \( \Delta \) is the mass of the market maker and \( Q^<_i \) is the random cumulative size of the orders queuing in front of market maker \( i \)’s order. Let all active market makers, other than \( i \), choose the same order size of \( q(h) \) (symmetric equilibrium). Hence, if market maker \( i \)’s initial order position turns out to be \( k_h \in [0, 1-h\delta/\eta] \), then \( Q^<_i = \int_0^{k_h} q(h) \, dj = k_h q(h) \). Further, by the continuum assumption, \( \Delta \) is infinitesimally small and the per capita expected profit upon execution becomes, in the limit,

\[ \lim_{\Delta \to 0} \frac{1}{\Delta} \mathbb{E}\left[\pi_h(Q^<_i + q_i\Delta) - \pi_h(Q^<_i)\right] = \mathbb{E}\pi_h(Q^<_i) q_i = \mathbb{E}\pi_h(K_h q(h)) q_i. \]

Using the above expression, the Bellman equation of market maker \( i \) can be written as

\[ u_h = \max_{q_i} \mathbb{P}(T \leq (h+1)\delta | T > h\delta) \mathbb{E}\pi_h(K_h q(h)) q_i \]

\[ + \mathbb{P}(T > (h+1)\delta | T > h\delta) \cdot \left[ \mathbb{P}\left(K_h \leq \min\left\{ k_h, \frac{\delta}{\eta}\right\}\right) \mathbb{E}\left(\pi_h(K_h q(h)) q_i \middle| K_h < \min\left\{ k_h, \frac{\delta}{\eta}\right\}\right) \right. \]

\[ + \mathbb{P}\left(K_h > \min\left\{ k_h, \frac{\delta}{\eta}\right\}\right) u_{h+1} \].

The first term on the right-hand side is the expected profit if the market order arrives before the next revision. The second term—if the market order arrives very late, allowing next the revision—has two components, depending on whether the market maker’s initial order will be queued to the front or to the rear. The cut-off point \( k^*_h \) determining “front” or “rear” is
defined, for \( h < \tilde{h} \), as

\[
(4.5) \quad k_h^*(q(h)) = \frac{1}{q(h)} \left( \bar{y} - \sum_{j=0}^{h-1} \delta \frac{q(j)}{\eta} \right).
\]

That is, if \( K_h > k_h^* \), the order queuing beyond \( \bar{y} \) and should be canceled.

A significant simplification is that \( q_i \) does not affect the continuation value \( u_{h+1} \) for all \( h \in \{0, 1, \ldots, \tilde{h}\} \), thanks to the continuum of market makers, each of whom has infinitesimally small (zero) influence on the total depth. Mathematically, \( \partial u_{h+1} / \partial q_i = 0 \). Thus, the first-order condition simplifies to

\[
\mathbb{P}(T \leq (h + 1)\delta \mid T > h\delta) \mathbb{E}\pi_h(K_h q(h)) + \mathbb{P}(T > (h + 1)\delta \mid T > h\delta) \mathbb{E}\left( K_h \leq \min\left\{ k_h^*, \frac{\delta}{\eta} \right\} \right) 
\cdot \mathbb{E}\left[ \pi_h(K_h q(h)) \mid K_h < \min\left\{ k_h^*, \frac{\delta}{\eta} \right\} \right] = 0
\]

and after evaluating the expectation terms (recall that \( K_h \) is uniformly distributed) and the probability terms, the first-order condition becomes

\[
(4.6) \quad \pi_h \left( \left( 1 - h \frac{\delta}{\eta} \right) q(h) \right) + \frac{1}{\lambda(h; \delta)} \pi_h \left( \min\left\{ k_h^*(q(h)), \frac{\delta}{\eta} \right\} q(h) \right) = 0,
\]

where \( \lambda(h; \delta) := (1 - G((h + 1)\delta))/(G((h + 1)\delta) - G(h\delta)) \) is the inverse (discrete) hazard rate of \( T \). The result is a system of difference equations which recursively updates \( \pi_h(y) \) according to equation (4.4) and \( k_h^* \) according to equation (4.5).

Equation (4.6) is a zero-profit condition for the continuum of (active) market makers. To see this, consider the baseline model of no revision opportunity; that is, \( \lambda(h; \delta) \to \infty \), annihilating the second term in equation (4.6). Observe further that the resulting first order condition is \( \mathbb{E}\pi_h(K_h q(h)) = \pi_h((1 - h\delta)q(h)) = 0 \), consistent with the baseline version as in equation (4.2). The only difference is that under the continuum assumption, the competition among market makers is perfect and hence the break-even condition is equivalent to the zero-profit condition.

The zero-profit condition (4.6) implies the following version of liquidity overshoot:

\[\text{Note that the left-hand side of the first-order condition (4.6) (i.e. } d_u/dq_i \text{) does not depend on market maker } i \text{'s order size } q_i \text{. Hence, as soon as this first-order condition holds, market maker } i \text{ is indifferent to any order size: As an infinitesimally small agent in the continuum, he has zero-mass and no influence on the aggregate liquidity supply. To close the equilibrium, the symmetric strategy assumption kicks in so that market maker } i \text{ also chooses } q_i = q(h).\]
Lemma 4.2 (Persistent liquidity overshoot). There is liquidity overshoot at all rounds of order revision until the order book stabilizes. Mathematically, $\pi(y(h)) < 0$ for $h \in \{0, \ldots, \bar{h} - 1\}$.

The results 1) and 2) claimed at the beginning of the equilibrium analysis can now be verified using lemma 4.2. For point 1): Suppose $\bar{h} = 0$, that is, there is no revision after an initial round of submission. Then, the initial submission must stabilize the book at $y(0) = 1 \cdot q(0) = \bar{y}$, contradicting the overshoot implication of equation (4.6). For point 2): For $h < \bar{h}$, a market maker whose initial order is no longer subject to queuing uncertainty at the $h$-th round is making positive expected profit (he knows that his order is queued before $\bar{y}$). He can, if desired, submit a new order or revise his existing order. However, given the time priority rule, a newly submitted order will queue at the end, implying expected loss because there is already liquidity overshoot. Revising down the size of the existing order only reduces the expected profit, while revising up the size moves the order to the end of the queue, losing in expectation. Hence, there is no incentive for the market makers with resolved queuing uncertainty to do anything.

It remains to determine $\bar{h}$, the last round revision. The recursion stops if there is no unresolved queuing uncertainty; that is, if the stabilized component in the true depth exceeds the break-even level: $\sum_{h=0}^{\bar{h}} \delta q(h) / \eta \geq \bar{y}$. Then the last revision always leads to $q(\bar{h}) = 0$ because any orders queued beyond $\bar{y}$ lose in expectation. Rearranging the inequality above using $q(\bar{h}) = 0$ gives the stopping criterion

\begin{equation}
\frac{\delta}{\eta} \geq \frac{1}{q(h-1)} \left( \bar{y} - \sum_{h=0}^{\bar{h}-2} \frac{\delta}{\eta} q(h) \right) = k_{\bar{h}-1}^*.
\end{equation}

That is, as soon as $k_{\bar{h}-1}^* \geq \delta / \eta$, the recursion stops in the next round, i.e. $\bar{h} = h + 1$. The following lemma summarizes the analysis above and gives a pure-strategy, symmetric equilibrium.

Lemma 4.3 (Equilibrium order submission and revision). At $t = 0$, all market makers submit a limit order of size $q(0)$ where $q(0)$ solves equation (4.6) with $h = 0$. At $t = h\delta$ for all $h \in \{1, \ldots, \bar{h} - 1\}$, if a market maker observes the queue position of his initial order to be $k < h\delta / \eta$, he does nothing; otherwise, he submits a revision order that changes his order size to $\min\{q(h-1), q(h)\}$ where $q(h)$ can be recursively solved from equations (4.4), (4.5), and (4.6). The recursion stops at $\bar{h}$ where $\bar{h} = \min\{h \mid h \geq 1, k_{h-1}^* \leq \delta / \eta\}$. 
4.4.3 Latency and order book depth dynamics

This subsection focuses on the effects of latency reduction on order book depth dynamics: Do lower latencies dampen or amplify liquidity overshoot? What are the implications on the stabilization of the order book? How do the effects of reaction latency $\delta$ and of transmission latency $\eta$ differ?

As seen in the previous subsection, a generic functional form of $G(\cdot)$ (which describes the arrival process of the market order) suffices to characterize the equilibrium book depth dynamics. In order to answer the above questions, however, the following parametrization is needed.

**Assumption 4.4 (Market order arrival).** The market order arrival time $T$ follows exponential distribution with $\mathbb{E}T = \tau > 0$. That is, $G(T) = 1 - e^{T/\tau}$ for all $T \geq 0$.

**Remark 1.** Exponential distribution is often used in modeling arrival time. In particular, the memorylessness property implies that the arrival probability in a fixed time interval is constant: $\mathbb{P}(T \leq (h + 1)\delta | T > h\delta) = 1 - e^{-\delta/\tau}$ for all $h \geq 0$. To this extent, the assumption is consistent with, for example, Foucault (1999) who assumes the trading process ends with the same probability in each period.

**Remark 2.** Note that assumption 4.4 implies $\partial G(T) / \partial \delta = \partial G(T) / \partial \eta = 0$. That is, changes in reaction or transmission latency does not affect the arrival probability of the market order. This would be the case if the market orders are submitted only by investors subject to shocks of information and liquidity needs, whose arrival processes are plausibly independent of trading technology changes. A more realistic setting should allow the market orders’ expected arrival time $\tau$ to be, for example, positively correlated with $\delta$. The positive correlation, however, hinders the tractability of the model and, yet, numerical procedures suggest that the results shown in this subsection remain qualitatively the same with or without a moderate correlation. Therefore, the independence of $\tau$ and the latencies $\delta$ and $\eta$ is maintained.

The magnitude of short-run liquidity overshoot depends on the initial limit order size $q(0)$. The equilibrium level of $q(0)$ should satisfy the zero-profit condition at $h = 0$:

$$\pi(q(0)) + \frac{1}{\tilde{\lambda}(\delta)} \pi\left(\min\left\{\tilde{y}, \frac{\delta}{\eta} q(0)\right\}\right) = 0.$$ 

(Note that $k_\eta(q(0)) = \tilde{y}/q(0)$.) In this equation, $q(0)$ is expressed as an implicit function of the latency parameters $\delta$ and $\eta$.

**Proposition 4.3 (Short-run liquidity overshoot and latency).** *Cateris paribus, short-run liquidity overshoot amplifies if either reaction latency or transmission latency reduces. That is, $\partial q(0) / \partial \delta \leq 0$ and $\partial q(0) / \partial \eta \leq 0$.***
CHAPTER 4. QUEUING UNCERTAINTY

(a) Varying reaction latency, $\delta$

(b) Varying transmission latency, $\eta$

Figure 4.4: Order book depth dynamics with varying latency. The left panel plots the depth dynamics for three specific levels of $\delta \in \{0.5, 0.2, 0.1\}$ with $\eta$ fixed at $\eta = 1$. The right panel plots the depth dynamics for three specific levels of $\eta \in \{1.2, 1.0, 0.7\}$ with $\delta$ fixed at $\delta = 0.15$. The numerical procedure is performed under the parametrization of $\pi(y) = y - y^2$ and $\tau = 1$.

The intuition behind the above result is based on the market makers’ improved order revision opportunity. With lower reaction latency, the market makers can react to the updates of market conditions at higher frequency. With lower transmission latency, the market makers observe more flow of information (better monitoring) in any fixed time interval. Therefore, the revision opportunity is better than in the high latency case, their expected profit is higher, and they will more fiercely compete and generate more pronounced liquidity overshoot.

Figure 4.4 illustrates the amplified short-run overshoot. Recall that the initial batch of orders (unit mass) is fully displayed by time $\eta$ (almost surely; see the definition of transmission latency). The peak at $t = \eta$ increases as $\delta$ decreases from 0.5 to 0.1 in panel (a) and as $\eta$ decreases from 1.2 to 0.7 in panel (b). (The numerical procedure and the parametrization are described at the end of this subsection.)

The improved revision opportunity implies that there will be more order revisions before the order book stabilizes. This raises the question whether the order book stabilization process slows down as the latencies reduce. To investigate this question, first note that at each but the last revision, the amount of orders submitted is $(1 - h\delta / \eta)$, which is the mass...
of market makers with unresolved queuing uncertainty. Second, the last batch of order amounts to mass \( (1 - (\bar{h} - 1)\delta/\eta) - k_{\bar{h}-1}^x \), which is the mass of market makers who realize that their initial orders are queued behind \( \bar{y} \) (hence canceling fully). Finally, recall that the transmission latency is defined so that it (almost surely) takes \( h \) units of time to display each unit mass of orders. Then define the stabilization time of the displayed order book by

\[
\bar{t}(\delta, \eta) := \eta \cdot \left[ \sum_{h=0}^{\bar{h}-1} \left( 1 - h \frac{\delta}{\eta} \right) + \left( 1 - (\bar{h} - 1) \frac{\delta}{\eta} \right) - k_{\bar{h}-1}^x \right]
\]

\[
= (\bar{h} + 1) \left( \eta - \frac{1}{2} \delta \bar{h} \right) + \left( \delta - \eta k_{\bar{h}-1}^x \right).
\]

Note that \( \bar{h} \) is implicitly defined by \( \delta \) and \( \eta \), and because \( \bar{h} \) can take only integer values, it turns out to be a step-function in the latency parameters. The direction of the step function can be signed:

| Lemma 4.4 (Number of revisions). | When the latencies are low, \( \bar{h} \) is a right-continuous, decreasing step function in \( \eta \) and is a left-continuous, increasing step function in \( \delta \). |

The lemma helps the derivation of the following proposition.

| Proposition 4.4 (Order book stabilization and latency). | In a low latency environment, cateris paribus, the stabilization of the order book shortens if transmission latency (\( \eta \)) lowers but prolongs if reaction latency (\( \delta \)) lowers. That is, \( \bar{t}(\delta, \eta) > \bar{t}(\delta, \eta') \) and \( \bar{t}(\delta, \eta) < \bar{t}(\delta', \eta) \) for \( 0 < \delta' < \delta \) and \( 0 < \eta' < \eta \). |

\textbf{Remark.} Comparative statics with respect to the latencies are nontrivial as the net effect depends on the indirect effects from all revision order sizes. Signing the comparative statics in general seems difficult. A key simplification to facilitate lemma 4.4 and proposition 4.4 results from the focus on low latency environment, in which \( \delta \) and \( \eta \) are small and the indirect effects are dominated by the direct effects. The details can be found in the proof.

The countervailing effects of transmission latency and reaction latency can be understood as follows: The former latency governs how fast the market events are played while the later measures how many such events are generated every moment. As the transmission speed increases (\( \eta \) drops; the movie being fast-forwarded), the stabilization occurs sooner than before. As more market events are being generated every moment (\( \delta \) drops; more frames inserted to the movie), naturally, it takes longer to have all these events processed and disseminated. The pattern is illustrated in figure 4.4.

The above proposition has important implications for empirical works. In particular, it emphasizes the difference between the two types of latencies. A technology improvement
that reduces the exchange’s latency perhaps should be distinguished from an innovation in
the traders’ trading technology. For example, consider an empiricist who obtains an order
book dataset that covers a period during which both reaction and transmission latency are
reduced. Intuitively, the two latencies should change in a correlated way. For example,
when the exchange provides colocation service (reducing transmission latency), it attracts
high-frequency traders who have ultra-low reaction latency. How does the speed upgrade
affect the book depth as seen from the empirical examination?

The answer depends on both the dominant force from the latency reductions and the
sampling frequency chosen by the empiricist. Consider figure 4.4. Fix an observation time
\( t \) and note that as the reaction latency drops in panel (a), the observed book depth increases
along the vertical slice. In panel (b), it can be seen that reducing the transmission does not
monotonically affect the displayed book depth at a fixed observation time \( t \).

Indeed, empirical evidences on how latency drops affect order book depth do not seem to
be unanimous. For example, on the transmission latency reductions, Riordan and Storken-
maier (2012) and Gai, Yao, and Ye (2013) find exchange speed upgrades lead to reduced
order book depth; on the other hand, Frino, Mollica, and Webb (2013) and Brogaard et al.
(2013) find introduction and upgrades of colocation service by exchange markets improve
observed book depth. The current model provides a plausible explanation for the “disagree-
ment”. It emphasizes that empirical works should construct depth (and other liquidity) vari-
ables with extra care when analyzing how latency affects market quality.

**Numerical illustration.** To illustrate the above comparative static results, the order book
depth dynamics are plotted in figure 4.4 with varying \( \delta \) and \( \eta \), respectively, under the fol-
lowing parametrization: \( \pi(y) = y - y^2/2 \) (implying \( \tilde{y} = 1 \)) and \( \tau = 1 \). For panel (a), \( \eta \)
is fixed at \( \eta = 1 \) and for panel (b), \( \delta \) is fixed at \( \delta = 0.15 \). The quadratic \( \pi(y) \) can be motivated
as a coarse approximation for the leading effects of other more realistic profit functions
of limit orders. Both propositions 4.3 and 4.4 are illustrated in the graph. In particular, as
transmission latency drops, the empirically examined order book depth does not necessarily
monotonically improve or worsen. Depending on the parametrization of \( \pi(\cdot) \), the resulting
shapes of the figures vary qualitatively.

**4.4.4 Policy discussion**

The above model-predicted order book depth dynamics captures some stylized empirical
facts, such as clustered order submission followed by immediate massive cancellation. Such
behavior is similar in appearance to the ill-purposed trading strategy known as “quote stuff-
ing”. For example, Egginton, van Ness, and van Ness (2012) refer to quote stuffing as “a
practice in which a large number of orders to buy or sell securities are placed and then
canceled almost immediately” and similar definitions are found in the media.\(^1\)

The model lends a rather innocuous explanation to the so-called quote stuffing behavior. The overshoot and the following immediate cancellation may simply be an equilibrium outcome due to queuing uncertainty, a market imperfection of low latency trading. This view echoes with Baruch and Glosten (2013)’s model prediction of “flickering quotes” in a zero-tick order book but with a different mechanism. In particular, the current paper focuses on the book depth dynamics, or “flickering depth”.

Along this line, this paper further argues that the recent regulation on limiting order-to-trade ratio and on minimum quote life is worth debating. For example, Oslo Børs has imposed a bound of 70 on the ratio of monthly orders to executed orders for each member, with an additional charge for exceeding this amount. Borsa Italiana has also implemented similar a order-to-trader policy in 2012. The London Stock Exchange imposed and has been adjusting a “high usage surcharge” on orders and quotes since 2010.

Such regulations limit market makers’ revision opportunities. In response, the market makers can be expected to reduce their order sizes in the short-run, leading to a thinner, less liquid order book. The MiFID II proposal rightfully expresses the concern that “to set out the maximum ratio of unexecuted orders to transactions” should take into account the liquidity of the financial instrument” (Article 51.7). The U.K. Foresight Projects also feature articles by Brogaard (2011), Friederich and Payne (2012), and Farmer and Skouras (2012), concerning similar negative effects on market quality. In particular, preliminary empirical evidence from Friederich and Payne (2012) finds that book depth in Italian stocks fell after an order-to-trade ratio limit is imposed.

4.5 Endogenous market order and welfare

The queuing uncertainty escalates the level of competition among market makers and hence pauperizes them. However, the overshoot benefits the liquidity demand side as the fundamental investors can trade to meet their liquidity needs at lower costs. This section adopts the fully endogenized framework to evaluate the overall effect of queuing uncertainty on social welfare. The analysis clarifies 1) how welfare depends on the liquidity in a limit order market and 2) under what conditions can the exchange improve welfare by randomizing the order queues.

Setup. The market settings are exactly the same as in Biais, Martimort, and Rochet (2000), except that queuing uncertainty is introduced and, for clarity, the focus is restricted to only one price level, \(a\), the best ask price. The one-price focus is motivated by the fact that in reality, there are few market orders that are large enough to bite into the book. This way, the

liquidity measure can be summarized in one scalar variable, the book depth. (With multiple
different prices, a complex “liquidity” measure is needed to account for variations in both price and
depth.)

There are two types of agents: One investor with CARA utility and \( n \geq 1 \) risk-neutral
market makers as described in section 4.3.1. The traded asset pays off \( V \) units of the
numéraire good upon consumption. The payoff \( V \) is defined as
\[
V = v_0 + S + e,
\]
where \( v_0 \) is
the unconditional expected payoff, \( S \) is the investor’s private signal with \( \mathbb{E}S = 0 \), and \( e \) is
the unobservable innovation, assumed to be normally distributed with mean zero. Denote
by \( \rho \) (\( > 0 \)) the product of the investor’s constant absolute risk-aversion coefficient and the
variance of \( e \). Timing of the game is as described in section 4.3.1.\(^1\) The investor observes
the signal \( S = s \) and also suffers from an endowment shock \( E = e \), where both realizations
\( s \) and \( e \) are her private information. Without loss of generality, let \( \mathbb{E}E = 0 \). The joint distri-
bution of \( E \) and \( S \) is commonly known. The stochastic queue of limit orders, the innovation
\( e \), and the joint distribution of \( E \) and \( S \) are all independent.

4.5.1 Equilibrium

Consider the investor’s optimization problem given a depth level \( y \) at the ask price \( a \). Con-
ditional on her private information of \( S = s \) and \( E = e \), the investor optimizes her certainty
equivalent:
\[
\max_{0 \leq x \leq y} m + (v_0 + \theta - a)x - \rho yx - \frac{\rho}{2}x^2
\]
where \( m := (s + v_0)(e + y) - \rho \cdot (e + y)^2/2 \) is the certainty equivalent for her endowment
and \( \theta := s - \rho e \) is her endowment-risk-adjusted signal for the asset. With the lower and the
upper bound of \( 0 \leq x \leq y \), the potentially cornered solution is
\[
x(\theta; y) = \min \left\{ y, \max \left\{ 0, \frac{1}{\rho} (v_0 + \theta - a) - y \right\} \right\}.
\]
It can be seen that the investor’s optimal order size truthfully reveals \( \theta \), an imperfect signal
(due to the liquidity/endowment shock; see also Vayanos and Wang, 2012), up to the lower
and the upper corners.

Consider next market makers’ profitability. Market makers do not know the investor’s
(endowment-risk-adjusted) signal, \( \Theta \), but they know \( \Theta := S - \rho E \), hence also its c.d.f.,
denoted by \( F(\theta) \) (the joint density of \( S \) and \( E \) is common knowledge). Suppose the depth
at this price \( a \) is \( y \geq 0 \). Then the expected profitability for these \( y \) units of the limit order is
\[
\pi(y) = \int_{a-v_0}^{a-v_0+\rho y} (a - v(\theta))x(\theta; y)dF(\theta) + \int_{a-v_0+\rho y}^{\infty} (a - v(\theta))ydF(\theta)
\]
\(^1\) To focus on the short-run order submission and to keep the model analytically tractable, unlike section
4.4, the game is only one-period as in Biays, Martimort, and Rochet (2000).
4.5. ENDOGENOUS MARKET ORDER AND WELFARE

where \( x(\theta; y) \) is the investor’s market order size at this price level (equation 4.8), and \( v(\theta) := \mathbb{E}[V|\Theta = \theta] \) is the conditional expectation of the asset payoff. Note that the integration begins from \( a - v_0 \) as \( x(\theta; y) \) is non-zero only if \( \theta > a - v_0 \), i.e. only if the signal is strong enough. Following Biais, Martimort, and Rochet (2000), two regularity conditions are needed. First, the joint distribution of \( S \) and \( E \) satisfies that \( v(\theta) \) is weakly increasing in \( \theta \), i.e. \( \dot{v}(\theta) \geq 0 \). Second, the support of \( \Theta \) is continuous.

The marginal expected profit is

\[
\pi(y) = \int_{a-v_0+\rho y}^{\infty} (a - v(\theta)) f(\theta) d\theta = \mathbb{P}(\Theta > a - v_0 + \rho y)(a - \mathbb{E}[v(\Theta)|\Theta > a - v_0 + \rho y]).
\]

(4.10)

The second equality gives the intuitive interpretation of the expected profit of the marginal order: Conditional on its execution (with probability \( \mathbb{P}(\Theta > a - v_0 + \rho y) \)), the sales revenue is the ask price \( a \) and the adverse-selection cost is \( \mathbb{E}[v(\Theta)|\Theta > a - v_0 + \rho y] \). The following proposition shows that “top-of-queue advantage” holds.

**Proposition 4.5 (Top-of-queue advantage).** There is top-of-queue advantage. In particular, both assumptions 4.1 and 4.2 hold for sufficiently large ask price, \( a \). Further, \( \pi(y) \) is quasi-convex in \( y \geq 0 \).

(A version of the top-of-queue advantage deep in the book is given later in section 4.6.) As soon as \( \pi(0) > 0 \), which is always true when the best ask price \( a \) is large, the analysis in the baseline model (section 4.3) applies, and the equilibrium order submission strategy is characterized by the first-order condition (4.2) and the existence is given by lemma 4.1.

### 4.5.2 Liquidity provision and welfare

Define welfare as the sum of the (ex ante) expected gains from trade of all participants, the investor and \( n \) market makers, measured by their respective ex ante certainty equivalents. Suppose the book depth is \( y \). The investor’s certainty equivalent is \( \text{ce}(y) = \mathbb{E}[M + (v_0 + \Theta - a)x(\Theta; y) - \rho x(\Theta; y)^2/2] \), where \( M := E \cdot (S + v_0) - \rho E^2/2 \) is the certainty equivalent of the investor’s endowment, \( \Theta := S - \rho E \) is the endowment-risk-adjusted signal, and \( x(\Theta; y) \) is the investor’s optimal market order size (which might be capped; see equation 4.14). The expected gains from trade is

\[
\text{ce}(y) - EM = \mathbb{E}[(\Theta - [\theta])x(\Theta; y) - \frac{\rho}{2} x(\Theta; y)^2],
\]

where \( [\theta] := a - v_0 \) is the floor which \( \Theta \) must exceed so that the investor will trade.

The market makers are risk-neutral and their aggregate expected certainty equivalent is just the sum of their expected profits.
Lemma 4.5 (Market makers’ aggregate profit). The aggregate profit of all market makers is queue-realization irrelevant. Only the aggregate book depth matters: \( \forall k \in \mathcal{K}, \sum_i \pi(Q_i^k(k) + q_i) = \pi(\sum_i q_i) \).

With lemma 4.5, the market makers’ aggregate expected gains from trade, fixing a book depth \( y \), can be computed as (c.f. equation 4.9)

\[
\pi(y) = \mathbb{E}[(\lfloor \theta \rfloor + v_0 - v(\Omega))x(\Omega; y)].
\]

Therefore, the social welfare, \((ce(y) - \mathbb{E}M) + \pi(y)\), is a function of the book depth, i.e. the liquidity level in this market:

\[
(4.11) \quad w(y) = \mathbb{E}[(v_0 + \theta)x(\Omega; y) - \frac{\rho}{2}x(\theta; y)^2 - v(\theta)x(\theta; y)].
\]

Note that the transfer, \( \lfloor \theta \rfloor x(\Omega; y) \), from the market makers (adverse selection cost) to the investor (information rent) offsets in the aggregation. The remaining terms have intuitive interpretations. The first term is the investor’s expected valuation of the position she buys from the market makers. The second term corrects her acquired risk. The last term is the value of the position seen from market makers’ perspective (recall that \( v(\theta) := \mathbb{E}[V|\Theta = \theta] \)).

Lemma 4.6 (Shape of welfare as a function of liquidity). For all \( y \geq 0 \), \( w(y) > \pi(y) \), \( \dot{w}(y) > \dot{\pi}(y) \), and \( \ddot{w}(y) < \ddot{\pi}(y) \). Further, \( w(y) \) is concavely monotone increasing on \( y \in [0, \bar{y}] \).

Lemma 4.6 only says that welfare is initially increasing. There might be, however, times when too much liquidity hurts social welfare. To see this, write the derivative of \( w(y) \) as

\[
\dot{w}(y) = \mathbb{P}(\Theta > [\theta] + \rho y)(\mathbb{E}[v_0 + \theta - \rho y|\Theta > [\theta] + \rho y] - \mathbb{E}[v(\theta)|\Theta > [\theta] + \rho y]).
\]

The difference between the two conditional expectations is the wedge between the investor’s and market makers’ valuation of the marginal unit of the asset. The difference is not necessarily positive as, from the social planner’s perspective, the investor might buy “too aggressively”. This is because though the investor enjoys her information rent, the social planner does not, as such rent offsets with the market makers’ adverse-selection cost. Consider the extreme case where the investor is very likely to draw a positive signal but no endowment shock: In this case, large liquidity provision in the market is socially suboptimal because it “indulges” the risk-averse agent to buy risky asset from the risk-neutral market makers. Put alternatively, the market makers’ competition, due to queuing uncertainty, might generate negative externality as the liquidity overshoot indulges inefficient reallocation of risky assets. Figure 4.5 shows an initially increasing welfare function eventually decreases if there is too much liquidity. The parametrization for the numerical procedure is described in section 4.5.4.
4.5. ENDOGENOUS MARKET ORDER AND WELFARE

Figure 4.5: Welfare as a function of liquidity supply. This figure qualitatively illustrates the shape of welfare as a function of aggregate liquidity supply (i.e. aggregate book depth). The numerical procedure is described in section 4.5.4. The vertical axis shows welfare, market makers’ expected profit, and the investor’s certainty equivalent. The horizontal axis shows the book depth, $y$. The upper curve (blue) is welfare, $w(y)$. The lower curve (red) is market makers’ aggregate expected profit, $\pi(y)$. The difference, indicated by the (green) vertical line, is the investor’s certainty equivalent, $ce(y)$.

Proposition 4.6 (Maximum social welfare). Social welfare maximizes at $y^*$, where $y^*$ is strictly larger than $\bar{y}$ (market makers’ break-even depth) and is possibly infinite.

Compare the cost-benefit analyses of a social planner and of a market maker:

$$w(y) - \pi(y) = \int_0^\infty [(v_0 + \theta - \rho y) - a]f(\theta)d\theta$$

$$= P(\Theta > |\theta| + \rho y)(E[v_0 + \Theta - \rho y|\Theta > |\theta| + \rho y] - a) > 0.$$  

The difference is exactly the investor’s expected benefit of buying a marginal unit of the asset. (If she could buy the marginal unit, she gains the conditional expectation term and she pays the ask price $a$.) That is, compared to a representative market maker who chooses the book depth to maximize the aggregate expected profit, a social planner in addition takes into account the investor’s additional gain. The socially optimal depth, therefore, always exceeds the market making optimal depth. The result is analogous to the scenario of a monopolist producer who, facing an exogenous demand function, chooses the production quantity to maximize his profit. There is always some consumer surplus cannot be realized because of the monopolist under-supplies in order to sell at a high price.

Recall that with queuing uncertainty, the order book deepens (liquidity overshoot) and hence there is room for welfare improvement by potentially adding queuing uncertainty to the market. This leads to the following subsection, which introduces queue randomization and analyzes how welfare, i.e. aggregate expected gains from trade, is affected.
4.5.3 Market design

In this subsection, consider $n = 2$ (representative) market makers, a fast one and a slow one. The two-agent assumption is an abstraction to reflect speed heterogeneity in the real-world trading environment, where some participants, having invested heavily in technology, have relatively low latency accessing the market compared to others. The two-agent restriction is imposed for the purpose of tractability. The numerical procedure (discussed in section 4.5.1) suggests that the results in this subsection hold generally.

Label the two market makers by 1 and 2, and there are two possible queues of the limit orders: $k \in \mathcal{X} = \{[1, 2]^T, [2, 1]^T\}$. Denote by $\alpha := \mathbb{P}(K = [1, 2]^T)$ the probability of market maker 1’s order queuing in the front of market maker 2’s. Without loss of generality, let market maker 1 be the (weakly) faster one: $\alpha \in [1/2, 1)$. In equilibrium, the limit order sizes, $q_1$ and $q_2$, should satisfy the following first-order condition system (c.f. equation 4.2):

\[
\begin{align*}
\alpha \pi'(q_1) + (1 - \alpha)\pi'(q_2 + q_1) &= 0, \quad \text{for market maker 1;} \\
(1 - \alpha)\pi'(q_2) + \alpha \pi'(q_1 + q_2) &= 0, \quad \text{for market maker 2.}
\end{align*}
\]

(4.12)

The following technical condition is needed for comparative statics:

\[
\pi(2\bar{y}) \leq 0.
\]

(4.13)

Then a connection between the aggregate book depth and speed heterogeneity can be established:

**Proposition 4.7** (Speed heterogeneity and liquidity overshoot). The liquidity overshoot is more pronounced when the market makers have homogeneous speed. Mathematically, if denote the equilibrium aggregate book depth by $y$, then $y(\alpha)$ has a local maximum at $\alpha = 1/2$.

**Corollary 4.1.** If $\pi(y)$ is “regular”, the equilibrium aggregate book depth $y$ is monotone increasing (decreasing) on $\alpha \in (0, 1/2)$ (on $\alpha \in (1/2, 1)$, respectively) and has its unique maximum at $\alpha = 1/2$.

**Remark.** A “regular” $\pi(\cdot)$ is important for the purpose of comparative statics. This is because the market makers’ problem, as alluded in proposition 4.2, resembles firms’ competition game and inherits the difficulty of signing comparative statics from such duo/oligopolistic competition games where agents’ decisions are strategic substitutes of each other. A convenient (sufficient) condition is that $\pi(\cdot)$ is concavely decreasing.

---

1 Condition (4.13) ensures that $\pi(\cdot)$ is concave on a support large enough. Effectively, it ensures the equilibrium is a “stable” one. The importance of a stable equilibrium for comparative statics has been emphasized in Samuelson (1941, 1942).
(c.f. a concavely decreasing demand function as in Kreps and Scheinkman, 1983) on the relevant support, as shown in the proof. When concavity does not hold, the proof shows that if the higher-order effects are small, this result is still valid.

“Speed” in the context is reflected in the probability of queuing in the front of the queue, and it is reflected by parameter $\alpha$ for market maker 1 and by $1 - \alpha$ for market maker 2. When $\alpha = 1 - \alpha = 1/2$, there is no speed heterogeneity. As $\alpha$ increases towards 1 (or decreases towards 0), the speed heterogeneity increases. By symmetry, the proposition implies that the overshoot reaches its maximum at $\alpha = 1/2$ and then monotonically decreases. Conceptually, therefore, less speed heterogeneity is equivalent to more queuing uncertainty, hence also more fierce competition among market makers. When there is no speed heterogeneity (most queuing uncertainty), the competition is most fierce and the most pronounced liquidity overshoot follows. When $\alpha \to 1$, market maker 1 is almost surely to be the first in the queue and will submit a limit order of size $\bar{y}$ to maximize his expected profit. The competition vanishes in this extreme case and there is no overshoot (setting $\alpha$ to 1 in the first-order condition system 4.12).

Now consider adjusting the queuing uncertainty in the market by introducing a queue randomizer, $R$, defined as a square random matrix

$$ R = \begin{bmatrix} B & 1 - B \\ 1 - B & B \end{bmatrix}, $$

where $B$ is a Bernoulli random variable with success probability $\beta \in [0, 1]$. For example, the exchange can apply the queue randomizer to every realized queue in a fixed time interval. Specifically, $R$ applies to a realized queue, $k$, by pre-multiplying it: The randomized queue becomes $K_R = RK$, which is either $K_R = k$ (with probability $\beta$) or $K_R \in \mathcal{K} \setminus \{k\}$ (with probability $1 - \beta$). Therefore, when the exchange adopts a queue randomizer $R$, the effective queue distribution can be characterized by

$$ \alpha_R(\beta) := \mathbb{P}(K_R = [1, 2]^T) = \alpha \beta + (1 - \alpha)(1 - \beta) \in [1 - \alpha, \alpha], $$

thanks to the simplified assumption that $n = 2$. When there are $n \geq 3$ market makers, the queue distribution is more complex to characterize. Appendix 4.8.A imposes additional structure to characterize the queue distribution. Section 4.5.4 adopts the micro-foundation in appendix 4.8.A to numerically illustrate the policy implication.

**Corollary 4.2 (Optimal queue randomizer).** There exists a randomizer $R$ such that welfare can be (weakly) improved when $R$ is applied.

Effectively, the queue randomizer adjusts the queuing uncertainty (equivalently, speed heterogeneity) and also the degree of competition between the market makers. For example,
consider the case where there is very severe speed heterogeneity, i.e. little queuing uncertainty. Without queue randomization, $y$ is close to $\bar{y}$ (lack of competition). By proposition 4.6, welfare can be improved by increasing the equilibrium book depth in this case. One way to do so is to add more queuing uncertainty by, for example, setting $\beta \leq 1/2$ so that the originally slow market makers are more likely to be re-queued to the front. This randomization curtails the “market power” of the fast market makers and strengthens the competition, which results in larger overshoot and a deeper book in equilibrium, benefiting the investors and improving welfare.

It should be noted, however, such queue randomization must be implemented with care for at least three reasons. First, the effectiveness of queue randomization is bounded by the initial speed heterogeneity among market makers. If there is little speed heterogeneity ($\alpha \sim 1/2$), the queue randomization is ineffective ($\alpha_R \sim 1/2$).

Second, improperly adjusting the queueing uncertainty can destroy social welfare. For example, suppose the unrandomized equilibrium aggregate book depth $y(\alpha)$ (with $\alpha \geq 1/2$) already exceeds the socially optimal depth $y^*$: $y^* < y(\alpha)$. In this case, improper queue randomization will reduce welfare because by adding more queuing uncertainty, the overshoot exacerbates, $y(\alpha_R(\beta)) \geq y(\alpha)$, and the increased liquidity supply is not appreciated by the social planner (see lemma 4.6 and its discussion).

Finally, by adding queuing uncertainty, the exchange essentially deprives the market power from fast market makers (who invest heavily in technology to achieve such speed advantage). The reduced profitability might, over time, constrain the participation of market makers. Such concerns, though beyond the scope of the current model, are legitimate and should be accounted for in choosing optimal queue randomization schemes.

### 4.5.4 Numerical example

This subsection adds normality to the model, following the example of Glosten (1994) and also example 1 of Back and Baruch (2013). Let the investor’s private signal and endowment shock be independent and both normally distributed with zero means. The endowment-risk adjusted signal is $\Theta = S - \rho E$, which is then also normally distributed with zero mean and variance $\text{var} \Theta := \text{var} S + \rho^2 \text{var} E$. The market maker’s inference, given a signal $\theta$, about the asset value is $v(\theta) = v_0 + \gamma \theta$, where the price impact factor is $0 < \gamma = \text{var} S / \text{var} \Theta < 1$. In the numerical illustration below, the parameters are $a = v_0 + 1$, $\text{var} S = 0.2$, $\text{var} E = 0.1$, and $\rho = 0.9$.

The investor’s certainty equivalent, the market makers’ aggregate expected profit, and welfare can then be evaluated using the expressions derived in section 4.5.2. Figure 4.5 illustrates the shape of welfare as a function of the aggregate book depth, contrasting with the market maker’s aggregate expected profit. It can be seen that welfare is not monotone increasing in liquidity supply in this example. Further, when market makers’ expected
4.5. ENDOGENOUS MARKET ORDER AND WELFARE

Figure 4.6: Queuing uncertainty and welfare. This figure illustrates how welfare and the equilibrium book depth respond to the level of queuing uncertainty in the market. The red curve is the equilibrium book depth ($y$, left axis) and the blue curve is welfare ($w$, right axis). The horizontal axis represents the level of queuing uncertainty, defined as the speed difference between market makers ($\Delta \mu$). The parameters are as described in section 4.5.4.

If profit maximizes, welfare can still be improved by increasing the liquidity supply from $\bar{y}$ ($\approx 0.2495$) to $y^*$ ($\approx 0.5216$).

When there is queuing uncertainty (and multiple market makers), the equilibrium book depth exceeds $\bar{y}$ and hence, there is room for welfare improvement by adding queuing uncertainty. However, when there is too much queuing uncertainty, the liquidity overshoot will be so excessive that welfare reduces. This would be the case when the equilibrium book depth exceeds $y^*$.

To visualize the effect directly, the following parametrization is prepared for figure 4.6. Let each market maker be endowed with a type $\mu_i$, such that the expected waiting time (latency) for market maker $i$’s order to be processed by the exchange server is $\mathbb{E}L_i = \mu_i$, where $L_i$ is exponentially distributed. The parameter $\mu_i$ is interpreted as the (expected) “latency” of market maker $i$; a faster market maker has a smaller $\mu_i$. Assume $L_i$ is independent of $L_j$ if $i \neq j$. Appendix 4.8.A gives the details and provides closed-form solutions to the queue distribution $\mathbb{P}(K)$. Finally, assume that the $n$ market makers all have heterogeneous speed: $\mu_i - \mu_{i+1} = \Delta \mu > 0$ for all $i \in \{1, \ldots, n-1\}$; that is, the speed difference is equally spaced. Note that, fixing $n$, increases in $\Delta \mu$ implies aggravated speed heterogeneity in market makers’ speed.\footnote{The parametrization adopted in this subsection for speed heterogeneity is one of the many possibilities. Other alternatives have been experimented and are shown to generate qualitatively similar patterns as shown in figures 4.5 and 4.6.}

Figure 4.6 plots the book depth (liquidity supply) and welfare in response of the level of queuing uncertainty in the market. The number of market makers is $n = 5$. The fastest
market maker’s speed is normalized to $\mu_1 = 1$ and $\Delta \mu$ ranges from 0 to 1.8. It can be seen that as queuing uncertainty increases, though the equilibrium book depth monotonically increases, welfare is first improved and then destroyed. The shape of the welfare curve is consistent with figure 4.5.

### 4.6 Deep in the book

This section studies the consequence of queuing uncertainty deep in the book. It shows the robustness of liquidity overshoot result (proposition 4.1), the driver of the order book depth dynamics in section 4.4 and the welfare analysis in section 4.5.

**Model setup with discrete prices.** The model extends the setup in section 4.5 so that multiple prices are possible. Unlike Biais, Martimort, and Rochet (2000, 2013), in order to make sense of queuing uncertainty, prices in the order book must be discrete. Let $\mathcal{P}$ denote the collection of all possible price levels in the limit order book: $\mathcal{P} = \{ p_{\min}, ..., p_{\max} \}$, where $-\infty < p_{\min} < v_0 < p_{\max} < \infty$. Describe the book depth by $y(p) : \mathcal{P} \mapsto \mathbb{R}$, which gives the (signed) depth at price level $p$. As before, because the analysis is symmetric, this paper only considers the ask side. Denote the $j$-th ask price by $a_j$ ($j \in \{1, 2, ...\}$) which is the $j$-th smallest price in $\{ p | p > v_0, p \in \mathcal{P}, y(p) > 0 \}$.$^1$ Let the depth at the $j$-th best price be denoted by $y_j$ and define $a_0 = v_0$ with $y_0 = 0$. (It will be justified that the ask side indeed corresponds to the prices $\{ p | p > v_0, p \in \mathcal{P} \}$ in equilibrium.) Finally, denote the cumulative depth up to and including (excluding) the $j$-th level by $y_j := \sum_{i \leq j} y_i$ (and $y_{\langle j} := \sum_{i < j} y_i$ similarly).

#### 4.6.1 The investor’s order size

Analogous to the analysis in section 4.5.1, consider the investor’s optimization problem given a book described by $y(p)$. Suppose the investor wants to buy the asset offered at the $j$-th ask price, with price $a_j$ and depth $y_j$ ($> 0$). Note that, the optimization is iterative: Due to price priority, the investor can only buy the offers at the $j$-th ask price if she has already bought everything offered below the $j$-th price. That is, she must have bought $y_{\langle j} units of the asset, accumulating her endowment to $e + y_{\langle j}$ units, before she could choose her optimal demand at price $a_j$. Then the investor optimizes:

$$\max_{0 \leq x_j \leq y_j} \mathbb{E} \left[ -\exp \left\{ -\left[ (E + y_{\langle j})V + (V - a_j) x_j \right] \right\} \right] | E = e, S = s]$$

---

$^1$Perhaps surprisingly, under queuing uncertainty, it is not necessary that the equilibrium ask prices are consecutive; i.e. it is possible that $a_{j+1} - a_j > \rho$. See corollary 4.3 for details.
4.6. DEEP IN THE BOOK

which is equivalent to maximizing her certainty equivalent:

$$\max_{0 \leq x_j \leq y_j} m_j + (v_0 + \theta - a_j)x_j - \rho y < j x_j - \frac{\rho}{2} x_j^2$$

where $m_j := (s + v_0)(e + y < j) - \rho (e + y < j)^2 / 2$ is the certainty equivalent for her holding (after buying everything offered below $a_j$) and $\theta := s - \rho e$ is her endowment-risk-adjusted signal for the asset. (Note that the transfer, $\sum_{j=0} y x_n$, that the investor has already paid does not come into the optimization.) The potentially cornered solution is

$$x_j(\theta; y_j) = \min\left\{y_j, \max\left\{0, \frac{1}{\rho} (v_0 + \theta - a_j) - y < j\right\}\right\}.$$  \hfill (4.14)

Clearly, equation (4.8) is a special case of the more general solution above. The corners give rise to the following “hurdles” that do not exist if the price is continuous:

**Lemma 4.7 (Hurdle in signal).** After the book depth at the $j$-th best price depletes, the investor further bites to the $(j+1)$-th price level if and only if her (endowment-risk-adjusted) private signal exceeds a strictly positive hurdle.

Mathematically, for each ask price $a_j$, there exist a floor $\lfloor \theta_j \rfloor = (a_j - v_0) + \rho y < j$ and a ceiling $\lceil \theta_j \rceil = (a_j - v_0) + \rho y < j$ such that $x_j(\theta)$ is off-corner if and only if $\lfloor \theta_j \rfloor < \theta \leq \lceil \theta_j \rceil$. The hurdle between two consecutive price levels is $\lfloor \theta_{j+1} \rfloor - \lceil \theta_j \rceil = a_{j+1} - a_j > 0$.

Figure 4.7 illustrates the investor’s optimal cumulative order size as a function of her endowment risk adjusted signal $\theta$ and the book $\{y_j\}$. This cumulative order size is essentially her demand for the asset (given the book depth), which is piece-wise linearly increasing with slope $1/\rho$ (the risk-adjustment).

4.6.2 Depths in the order book

Consider the $j$-th ask price level at $a_j = p > v_0$, with the previous cumulative depth $y < j$. Suppose the depth at this price $a_j$ is $y \geq 0$ and then the expected profitability for these $y$ units of the limit order is

$$\pi_j(y) = \int_{\lfloor \theta_j \rfloor + \rho y}^{\lceil \theta_j \rceil + \rho y} (a_j - v(\theta)) x_j(\theta) f(\theta) d\theta + \int_{\lfloor \theta_j \rfloor + \rho y}^{\infty} (a_j - v(\theta)) y f(\theta) d\theta$$  \hfill (4.15)

where the signal floor $\lfloor \theta_j \rfloor$ is defined in lemma 4.7, $x_j(\theta)$ is the investor’s market order size at this price level (equation 4.14), $v(\theta) := E[V \mid \Theta = \theta]$ is the conditional expectation of the asset payoff, and $f(\theta)$ is the density of $\Theta$ (the density is assumed to exist, following Biais, Martimort, and Rochet, 2000). The marginal expected profit is

$$\bar{\pi}_j(y) = \int_{\lfloor \theta_j \rfloor + \rho y}^{\infty} (a_j - v(\theta)) f(\theta) d\theta.$$  \hfill (4.16)
The investor’s optimal market order size. The investor’s cumulative order size, \( x_{\infty} := \sum_j x_j \), is plotted against her endowment-risk-adjusted signal, \( \theta \), given a limit order book with depths \( y_j \) at the \( j \)-th best ask price. The curve is piece-wise increasing in \( \theta \). The investor bites into price level \( a_j \) only if \( \theta \) exceeds a floor threshold \( \lfloor \theta_j \rfloor \) and depletes that level only if \( \theta \) exceeds a ceiling \( \lceil \theta_j \rceil \) as given in lemma 4.7. The hurdle size is \( a_{j+1} - a_j \).

The profit and marginal profit expressions (4.15) and (4.16) generalize the one-price level versions in equations (4.9) and (4.10), respectively. The following is a generalized version of proposition 4.5 seen in section 4.5.1.

**Proposition (Top-of-queue advantage).** There is top-of-queue advantage in the limit orders at each price level. In particular, both assumptions 4.1 and 4.2 hold. Further, \( \pi_j(y) \) is quasi-convex in \( y \geq 0 \).

Now turn to the market makers. Each market maker submits a set of limit orders, one order for each price level, to maximize his expected profit. Because the cumulative depth at any price level affects the profitability of orders deep in the book, the optimization is dynamic (along the price dimension) in nature. Consider market maker \( i \)'s order size at price level \( j \), given all other market makers’ order sizes. Denote by \( u_{i,j} \) his continuation value for price level \( j \) and the Bellman equation can be written as

\[
(4.17) \quad u_{i,j} = \max_{q_{i,j}} \mathbb{E} \left[ \pi_j \left( Q_{i,j}^\infty + q_{i,j} \right) - \pi_j \left( Q_{i,j}^\infty \right) \right] + u_{i,j+1},
\]

where for notation simplicity, the arguments of \( u_{i,j+1} \) are omitted but it should be understood that it depends on market maker \( i \)'s order size \( q_{i,j} \) at price level \( j \). With some manipulation, it can be shown that the first-order condition can be written as (see the proof of proposition 4.8 in appendix)

\[
(4.18) \quad \mathbb{E} \pi_j \left( Q_{i,j}^\infty + q_{i,j} \right) - \mathbb{E} \pi_{j+1} \left( Q_{i,j+1}^\infty \right) = 0.
\]
Compared with the first-order condition in the baseline model (equation 4.2), there appears a new term, \(-\mathbb{E} \hat{\pi}_{j+1} \left( Q_{i,j+1}^c \right)\), in the equation. This new term measures the expected profit of the first marginal unit of the order on price level \(j+1\). It is negative because if market maker \(i\) increases his order size at level \(j\), the profitability at level \(j+1\) reduces (lower execution probability and higher adverse-selection cost). The interpretation of this first-order condition is, instead of myopically choosing the order size to break even the expected profit at the current price level, market maker \(i\) also accounts for his expected profitability on the next price level. This effect, however, is not large enough to curb the fierce competition among market makers and the aggregate liquidity at each price level still overshoots, as proved in the following proposition.

**Proposition 4.8 (Overshoot deep in the book).** Suppose the investors’ endowment-risk-adjusted signal \(\Theta\) exhibits (weakly) increasing hazard rate. Then, in equilibrium, there is liquidity overshoot at each price level. Mathematically, \(\hat{\pi}_j(\sum_i q_{i,j}) < 0\) for all \(j\) and the equality holds if and only if there is one market maker who is almost surely the first in queue at price level \(j\).

**Remark.** The technical condition that \(\Theta\) exhibits (weakly) increasing hazard rate, i.e., \(f(\theta)/(1 - F(\theta))\) is weakly increasing in \(\theta\), is a sufficient (but not necessary) condition for the overshoot result. As discussed in Biais, Martimort, and Rochet (2000), who also assume such a condition (equations 16 and 17), this is not a restrictive assumption. Some commonly used distributions (to name a few, uniform, normal, and exponential) all satisfy increasing hazard rate. The hazard rate essentially reflects the severity of market makers’ adverse-selection problem. Therefore, the increasing hazard rate assumption can also be understood, intuitively, as a condition to ensure the regularity of the problem; see Baruch and Glosten (2013) for more details on the importance of sufficient adverse-selection in the limit order market.

The overshoot results suggests that instead of break-even, the last unit limit order at each price level loses in expectation, at least in very short run. This gives an equilibrium prediction about “holes” in the limit order book (in short-run):

**Corollary 4.3 (Holes in book).** For sufficiently small tick size, the equilibrium limit order book has, possibly interior, “holes”, or empty price levels on which no market makers are willing to post any orders.

To understand the intuition of this result, it is helpful to look at again the expression of the expected marginal profit (equation 4.18). When the price level goes from \(j\) to \(j+1\), the (expected) marginal revenue jumps from \(a_j\) to \(a_{j+1}\) and, yet, the expected marginal cost, which only depends on the cumulative book depth, does not change. Liquidity overshoot implies that the expected marginal cost of the last limit order at price level \(j\) exceeds \(a_j\).
Hence, if the next price level \( a_j + \rho \) (where \( \rho \) is the tick size) is not large enough, the marginal profit at price \( a_j + \rho \) will never be positive and no market maker will place order there.

Empirical evidence of holes in limit order book has been found by, for example, Biais, Hillion, and Spatt (1995). Queuing uncertainty gives an equilibrium explanation. In particular, equilibrium holes are not supported by the equilibrium of Sandås (2001), whose break-even condition asserts that the last unit of limit orders at each price level has expected revenue equating expected cost. This is because, as the price level moves to the next tick, the expected revenue jumps up but the expected cost does not change, implying strictly positive expected marginal profit for the first unit of limit orders at the new price. Market makers will compete for such profit by submitting orders to the new level, and there will be no holes.

It should be emphasized that the predicted holes in the book are *transitory* in nature. If the market makers are given sufficient time to revise/cancel their orders, the long-run equilibrium reverts to the break-even level described by Sandås (2001) and the holes will disappear; see section 4.4 for the book depth dynamics.

A second corollary follows proposition 4.8. It compares the cumulative liquidity in the book under queuing uncertainty and under the break-even condition of Sandås (2001).

**Corollary 4.4 (Cumulative liquidity overshoot).** Denote by \( y^u_{\leq j} \) and \( y^b_{\leq j} \), respectively, the equilibrium cumulative depths until (inclusive) the \( j \)-th price level under queuing uncertainty and under the break-even condition (Sandås, 2001). If there are no holes in the book under queuing uncertainty, then \( y^u_{\leq j} \geq y^b_{\leq j} \) for all price levels.

This corollary is not a restatement of proposition 4.8 because the proposition focuses on the depth at each price level while the corollary addresses the *cumulative* depth. In particular, the overshoot at price level \( j \) affects the break-even quantity at price level \( j + 1 \).

### 4.6.3 Alternative setups

So far this section has been building on the fully-endogenized framework of Biais, Martin-mort, and Rochet (2000). This subsection gives two other examples, following Back and Baruch (2013), to show that the top-of-queue advantage (assumptions 4.1 and 4.2) holds generically.

**Example 4.1** (Informed or liquidity dichotomy). This example adopts the classical dichotomy that treats the incoming market order to be driven by either private information or liquidity shock. Such dichotomy is seen in, for example, Glosten and Milgrom (1985), Kyle (1985), Easley and O’Hara (1992), and many others.
4.7. Conclusion

Suppose that with probability \( \psi \) the arriving investor observes the true value of the asset, \( v \), while with probability \( 1 - \psi \) the arriving investor is purely liquidity driven. If the investor is information driven, she buys all units of the asset offered at prices \( p \leq v \) (again, the example only deals with the ask side). If the investor is liquidity driven, she submits a market order of size \( X \), which has density \( f(x) \) defined on \( \mathbb{R} \). Then market makers’ expected profit of a limit order of size \( y \) at price level \( a_j \) is

\[
\pi_j(y) = \psi \cdot \left( a_j - \mathbb{E}[V | V \geq a_j] \right) y + (1 - \psi) (a_j - v_0) \left[ \int_{y_j}^{y_j+y} (x - y_j) f(x) \, dx + \int_{y_j+y}^{\infty} y f(x) \, dx \right]
\]

where, following the notations developed in this section, \( y_j \) is the cumulative depth excluding the \( j \)-th price level. It is easy to evaluate that \( \pi_j(y) = -(1 - \phi) (a_j - v_0) f(y_j + y) < 0 \); that is, \( \pi_j(y) \) is strictly concave at any price level \( a_j > v_0 \). Therefore, assumption 4.1 holds under this setup, with the quasi-concavity strengthened to concavity, and hence follows assumption 4.2.

**Example 4.2** (Inelastic demand). In this example, the incoming market order size follows an exogenous distribution that is independent of the order book depth. Such models are seen in, for example, Sandås (2001), Foucault and Menkveld (2008), and van Kervel (2013).

Denote the market order size by \( X \), a random variable with density function \( f(x) \). Denote the price impact by \( v(x) := \mathbb{E}[V | X = x] \) and assume \( v(x) \geq 0 \). Then given the cumulative depth \( y_j \), market makers’ expected profit of a limit order of size \( y \) at price \( a_j \) is

\[
\pi_j(y) = \int_{y_j}^{y_j+y} (a_j - v(x)) (x - y_j) f(x) \, dx + \int_{y_j+y}^{\infty} (a_j - v(x)) y f(x) \, dx.
\]

The expression is akin to equation (4.15), the fully-endogenized version. Indeed, following the proof of proposition 4.5, it can be shown that \( \pi_j(y) \) is quasi-concave and, further, \( \pi_j(y) > 0 \) implies \( \pi_j(y) < 0 \) as in proposition 4.5, satisfying assumptions 4.1 and 4.2.

The other results derived in section 4.6.2 (proposition 4.8 and its two corollaries) are still valid under the setups of these two examples. The derivation and proofs are omitted for brevity.

**4.7 Conclusion**

In a high-speed trading environment, traders cannot perfectly condition their decisions on real-time market status. This is because between a market status update and a trader’s registering of the update, there is non-zero delay during which the market status is likely to
evolve. This paper acknowledges such market imperfection by introducing queuing uncertainty to standard limit order market model. The agents strategically play a simultaneous game that results in liquidity overshoot in equilibrium because the queuing uncertainty intensifies the competition and weakens the strategic substitution effects of others’ decisions.

The model identifies transitory overshoot in liquidity provision as an equilibrium phenomenon in modern financial markets. Further, it is argued that flickering orders and the so-called “ghost liquidity” naturally follow the overshoot of liquidity as liquidity providers cancel and revise those orders that end up in the bottom of the queue. The transitory nature of the liquidity overshoot is associated with two different types of latencies in the limit order market: reaction latency and transmission latency. Although reduction in either type of the latencies amplifies the transitory liquidity overshoot (because faster speed accessing the market guarantees better order revision opportunity and hence intensifies competition), the implications on the stabilization process of the limit order book are different. While drops in reaction latency prolong the stabilization (increased amount of revisions in a given time interval), drops in transmission latency hasten the stabilization (reduced update time for a given amount of revisions). The model predictions concur with recent empirical literature on how latency drops affect order book depth.

The model also adds to the recent debate on trading speed and related regulations. After endogenizing liquidity demanders (fundamental investors), the model examines market quality through the expected gains from trade of all agents in the economy and shows that the exchange can improve market quality by cautiously randomizing the order queues in the limit order book, effectively adjusting the level of queuing uncertainty. Caveats on the implementation of queue randomization are discussed.

This paper invites future empirical works to test the model’s prediction about order book depth dynamics and to verify the source of liquidity improvement due to the advancement in trading technology. The theoretical framework developed in this paper also leaves room for further exploration of how particular regulations (for example, make/take fees, minimum order resting time, etc.) may affect traders’ order submission and market quality under queuing uncertainty.
4.8 Appendix

4.8.A Speed and queuing

This appendix models how the stochastic queue, \( K \), is realized according to market makers’ speed. The probability masses of potential \( n! \) queuing outcomes are fully characterized by \( n \) speed parameters of the market makers. The model specification captures a key feature that speed is \textit{relative} in nature: Increasing all market makers’ speed by a same factor should not affect the distribution of the queue.

Endow each market maker \( i \) with a speed \( \mu_i^{-1} > 0 \). In what follows, \( \mu \) will be interpreted as the (expected) latency and, interchangeably, its inverse, \( \mu^{-1} \), will be referred to as the speed. Collate these speed parameters in a vector \( \mu = [\mu_1, \ldots, \mu_n]^\top \). Let Nature draw for each market maker a latency, \( L_i (> 0) \), from a commonly known distribution with c.d.f. \( G(l_i; \mu_i) \) where \( l_i \) is defined on \((0, \infty)\). The latency draws are assumed to be independent. Order the realized latencies as \( l_{i(1)} \leq \cdots \leq l_{i(n)} \), where \( i(k) \) is the index of the market maker who draws the \( k \)-th smallest latency. Each market maker \( j \)’s queue position then realizes to be \( K_j = k \) if \( i(k) = j \).

The latency (or, simply, delay) measures how soon the market maker can “effectuate” the order size she chose. Such effectuation delay, in a limit order market, can be the waiting time for the order to travel from a market maker’s computer to the exchange server.

**Assumption 4.5 (Latency distribution).** Latency \( L_i \) is exponentially distributed on \((0, \infty)\) with mean \( \mu_i (> 0) \). That is, \( \mathbb{P}(L_i \leq l) = 1 - e^{-l/\mu_i} \) for \( l > 0 \). \( L_i \) is independent of \( L_j \) for all \( j \neq i \).

It is acknowledged that this assumption is not without loss of generality. First, it rules out simultaneity almost surely (\( \mathbb{P}(L_i = L_j) = 0 \) if \( i \neq j \)). Second, all market makers have strictly positive probability to be in any queue position, i.e. \( \mathbb{P}(K_i = k) > 0 \) for all \( i \) and \( k \). This rules out, for example, the case where market maker 1 is never the first. In exchange, assumption 4.5 brings tractability and the intuitive property that only relative speed matters for the distribution of the queue.

Probability mass function for the queue vector \( K \) can be easily solved. Consider a queue realization \( k = \{k_1, \ldots, k_n\} \) determined by latencies \( L_{i(1)} < \cdots < L_{i(n)} \). The probability of realizing such a queue is

\[
\mathbb{P}_\mu(K = k) = \mathbb{P}_\mu(L_{i(1)} < \cdots < L_{i(n)}) \\
= \int_0^{\infty} \cdots \int_0^{\infty} \mathrm{d}G(l_{i(1)}; \mu_{i(1)}) \cdots \mathrm{d}G(l_{i(n)}; \mu_{i(n)}),
\]
CHAPTER 4. QUEUING UNCERTAINTY

which under assumption 4.5 has closed-form solution

\[(4.19) \quad \mathbb{P}_\mu(K = k) = \prod_{j=1}^{n} \frac{\mu_i^{-1}(j)}{\sum_{h=j}^{n} \mu_i^{-1}(h)}.\]

The subscript \( \mu \) on \( \mathbb{P}(\cdot) \) emphasizes the dependence of the probability measure \( \mathbb{P} \) on market makers’ speed \( \mu \). The following two lemmas will be useful in characterizing the queue distribution.

**Lemma 4.8 (Speed relativity).** Under assumption 4.5, only market makers’ relative speed matters for the queue distribution. Mathematically, \( \mathbb{P}_\mu(K) = \mathbb{P}_{c\mu}(K) \) for any \( c > 0 \).

**Lemma 4.9 (Speed increase and queue distribution).** Holding all other market makers’ speed constant, an increase in the speed of a market maker pushes him to the front of the queue and all others to the back (in a probability sense).

Mathematically, write \( C_i \) as a diagonal matrix of size \( n \) with the \( i \)-th diagonal term equal to \( c > 1 \) and all other diagonal terms equal to 1. Then, for any \( i \neq j \), the distribution \( \mathbb{P}_\mu(K_i) \) first-order stochastically dominates the distribution \( \mathbb{P}_{c\mu}(K_i) \) and the distribution \( \mathbb{P}_\mu(K_j) \) is first-order stochastically dominated by the distribution \( \mathbb{P}_{c\mu}(K_j) \).

Rather intuitively, lemma 4.8 says if all market makers’ speed are boosted by a same (strictly positive) factor, the resulting distribution of their queue does not change. That is, market makers’ latencies are ordinal. Lemma 4.9 qualifies the effect on queue distribution of an increase in market maker \( i \)’s speed. The vague idea that faster market makers are more likely to queue in the front is proved, under assumption 4.5, in terms of stochastic dominance.

### 4.8.B Notation summary

**General notations:**
- \( N \), the set of all market maker indices, \( N = \{1, \ldots, n\} \).
- \( \pi \) (also \( \hat{\pi}, \tilde{\pi}, \) etc.): the aggregate expected profit (and its derivatives) of all limit orders.
- \( q_i \): the limit order size of market maker \( i \).
- \( q \): the vector collecting all market makers limit order sizes, \( q = [q_1, \ldots, q_n]^T \) and \( q_{-1} = q \setminus \{q_i\} \).
- \( k \): queue position index.
- \( k \): queue realization, a vector of size \( n \).
- \( \mathcal{K} \): the collection of all possible queues realizations.
- \( n \): the total number of market makers.
- \( i \): market maker index, \( i \in \{1, \ldots, n\} \) (in section 4.4, \( i \in [0, 1] \) under assumption 4.3).
• $Q^c_i(k)$: the aggregate depth before (excluding) market maker $i$’s order, given the queue realization $k$.
• $y$: the book depth.
• $\bar{y}$, the “break-even” book depth level at which $\pi(\bar{y}) = 0$.

Additional notations in section 4.4:
• $\delta$: market makers’ reaction latency.
• $\eta$: transmission latency.
• $h$: the number of order revisions.
• $\bar{h}$: the maximum number of order revisions.
• $\lambda(h; \delta)$: the inverse (discrete) hazard rate of $T$, defined as $(1 - G((h + 1)\delta))/(G((h + 1)\delta) - G(h\delta))$.
• $q(h)$: the after-revision order size for the active market makers.
• $\bar{r}$: the time at which the displayed order book stabilizes.
• $T$: the (exponentially distributed) stochastic arriving time of the market order. It has c.d.f. $G(T)$ with support $[0, \infty)$.
• $y(h)$: the true book depth after the $h$-th revision.

Additional notations in section 4.5:
• $a$: the ask price.
• $\alpha$: the probability of market maker 1 being first in queue; only in section 4.5.3.
• $B$: a Bernoulli random variable with success probability $\beta$; only in section 4.5.3.
• $ce$: the investor’s (ex ante) certainty equivalent.

Additional notations in section 4.6:
• $a_j$: the $j$-th best ask price; $a_j \in \{p | p > v_0, p \in \mathcal{P}\}$.
• $j$: the $j$-th best (ask) price level.
• $\mathcal{P}$: the collection of all possible price levels in the order book with min $\mathcal{P} = p_{\min} < \max \mathcal{P} = p_{\max}$.
• \([\theta_j]\) and \([\theta_j]\): the floor and the cap that bind the investor’s market order size at price level \(j\) (see lemma 4.7).
• \(x_j\): the investor’s market order size (net demand) at price level \(j\).
• \(y_{\leq j}\) (and \(y_{<j}\), respectively): the cumulative depth up to and including (excluding) the \(j\)-th level.

4.8.C Proofs

Lemma 4.1

Proof. First, note that by quasi-concavity of \(\pi(\cdot)\) and the terminal conditions, there exists a \(\bar{y} > 0\) such that \(\pi(\bar{y}) = 0\), i.e. the \(\bar{y}\)-th marginal order breaks even (in expectation). Second, no market maker will submit a limit order larger than \(\bar{y}\) because the part exceeding \(\bar{y}\) always loses (negative marginal profit). By reducing the order size to \(\bar{y}\) strictly improves his total profit. This reduces each market maker’s strategy space to \([0, \bar{y}]\), and the best response correspondences implied by the first-order conditions can be summarized as a vector-valued function \(f: [0, \bar{y}]^n \rightarrow [0, \bar{y}]^n\), a nonempty, compact, and convex set. Note that the convexity of the value set of \(f\) is implied by the continuity of the first-order conditions (\(\pi(\cdot)\) is twice differentiable by assumption 4.1). Then by Kakutani’s fixed point theorem, there exists at least one fixed point to the first-order condition system.

It remains to show that among the fixed points, at least one satisfies the second-order condition. To see this, fix \(q_{-i}\) and consider market maker \(i\)’s first-order condition as derived in equation (4.2). Clearly, by construction \(\mathbb{E}\pi(Q_i^<(K) + \bar{y}) \leq 0\). If \(\mathbb{E}\pi(Q_i^<(K) + q_i) \leq 0\) for all \(q_i \in [0, \bar{y}]\), then this market maker will not participate (he is too slow and drops out). Otherwise, by continuity, there exists at least one \(q_i \in [0, \bar{y}]\) such that for the given \(q_{-i}\), both the first-order and the second-order conditions hold. Applying the same argument to all market makers yields at least one fixed point that is indeed an equilibrium.

Finally, \(q = 0\) is not an equilibrium because at least one market maker wants to deviate (\(\pi(0) > 0\), assumption 4.1). Therefore, at least one \(q_i\) is strictly positive in equilibrium. \(\square\)

Lemma 4.2

Proof. Rearranging the zero-profit condition 4.6 to get

\[
\pi(y(h)) = -\lambda(h; \delta)^{-1}\pi_h \left( \min \left\{ k_h^*, \frac{\delta}{\eta} \right\} q(h) \right) < 0,
\]

where \(y(h)\) is the true book depth defined in equation (4.3). The inequality follows due to the construction of \(k_h^*\). This implies, by assumption 4.1, that \(y(h) > \bar{y}\), i.e. there is liquidity.
overshoot. The above inequality holds for all $0 \leq h < \tilde{h}$. At $h = \tilde{h}$, by definition, the book stabilizes and hence there is no more liquidity overshoot.

\section*{Lemma 4.3}

\emph{Proof.} The lemma directly follows the analysis in section 4.4.2.

\section*{Lemma 4.4}

\emph{Proof.} First, fix $\delta$ and examine changes in $\eta$. Denote by $\bar{y}(h)$ the stabilized part of the true depth (see equation 4.3) before the $h$-th revision. Then $\bar{y}(h) = \sum_{j=0}^{h-1} \delta q(j)/\eta$, and

$$ \frac{\partial \bar{y}(h)}{\partial \eta} = \frac{\delta}{\eta^2} \left[ \sum_{j=0}^{h-1} q(j) - \eta \sum_{j=0}^{h-1} \frac{\partial q(j)}{\partial \eta} \right] \approx -\frac{1}{\eta} \bar{y}(h) \leq 0, $$

where the approximation is valid for small $\eta$. Hence, as the transmission latency reduces, the stabilized book depth before each revision (weakly) increases. Suppose $\tilde{h} = h$. Then $\bar{y}(h) \geq \bar{y}$ but $\bar{y}(h-1) < \bar{y}$ and as $\eta$ drops, $\bar{y}(h-1)$ keeps increasing until it reaches $\bar{y}(h-1) = \bar{y}$. At $\bar{y}(h-1) = \bar{y}$, after the stopping criterion (4.7) is satisfied by $\tilde{h} = h - 1$. Hence, $\tilde{h}$ is an increasing right-continuous function in $\eta$. Note that at $\eta \to 0$, $\tilde{h}$ reaches its minimum at $\tilde{h} = 1$.

Similarly, fix $\eta$ and examine changes in $\delta$ next.

$$ \frac{\partial \bar{y}(h)}{\partial \delta} = \sum_{j=0}^{h-1} \frac{1}{\eta} q(j) + \bar{y} \sum_{j=0}^{h-1} \frac{1}{\eta} \frac{\partial q(j)}{\partial \delta} \approx \frac{1}{\delta} \bar{y}(h) \geq 0, $$

where the approximation is valid for small $\delta$. Hence, as the reaction latency increases, the stabilized book depth before each revision (weakly) increases, and in particular, if $\tilde{h} = h$, then $\bar{y}(h-1)$ keeps increasing, until it reaches $\bar{y}(h-1) = \bar{y}$, satisfying the stopping criterion (4.7) at $\tilde{h} = h - 1$. That is, $\tilde{h}$ is a decreasing left-continuous function in $\delta$. Note that if $\delta \to 0$, $\tilde{h} \to \infty$.

\section*{Lemma 4.5}

\emph{Proof.} Fix a limit order vector $q$ and a possible queue realization $k$. The expected profit of market maker $i$, given the queue, is $\pi(Q_i^c(k) + q_i) - \pi(Q_i^c(k))$. Sum over all market makers’ expected profit to get

$$ \sum_i [\pi(Q_i^c(k) + q_i) - \pi(Q_i^c(k))] = [\pi(q_{i(1)}) - \pi(0)] + [\pi(q_{i(1)} + q_{i(2)}) - \pi(q_{i(1)})] + \cdots + [\pi\left(\sum_{k=1}^{n} q_{i(k)}\right) - \pi\left(\sum_{k=1}^{n-1} q_{i(k-1)}\right)] = \pi\left(\sum_i q_i\right). $$
In the equation, subscripts on $q$ means that $i(k)$ is the index of the agent at the $k$-th position of the queue. The last equality follows because all the in-between $\pi(\cdot)$ terms cancel. (Recall also that $\pi(0) = 0$.) The aggregate expected profit is queue-irrelevant.

**Lemma 4.6**

**Proof.** That $w(y) > \pi(y)$ is trivial because by definition welfare accounts for both the certainty equivalent of the investor and the expected profit of market makers: $w(y) = \pi(y) + ce(y)$. Evaluate the difference between $\dot{w}(y)$ and $\dot{\pi}(y)$ to get

$$\dot{w}(y) - \dot{\pi}(y) = \int_{[\theta]+\rho y}^{\infty} (v_0 + \theta - \rho y - a)f(\theta)d\theta > 0.$$  

The integration is positive because $q > \lfloor q \rfloor + r y = a + v_0 + r y$. Therefore, for all $y \geq 0$, $\dot{w}(y) > \dot{\pi}(y)$. By proposition 4.5, $\pi(y)$ exhibits quasi-concavity and hence, for all $0 \leq y \leq \bar{y}$, $\dot{w}(y) > \dot{\pi}(y) \geq 0$. Similarly, after some simplification, $\ddot{w}(y) - \ddot{\pi}(y) = -\rho \cdot (1 - F(\lfloor q \rfloor + r y)) < 0$ for all $y$. By proposition 4.5, $\ddot{w}(y) < \ddot{\pi}(y) < 0$ on $y \in [0,\bar{y}]$. Hence, $w(y)$ is concavely increasing on $[0,\bar{y}]$. 

**Lemma 4.7**

**Proof.** Solving the lower and the upper corners of equation (4.14), i.e.

$$\begin{cases} x_j(\lfloor \theta_j \rfloor) = 0 \\ x_j(\lceil \theta_j \rceil) = y_1 \end{cases}$$

gives the floor and the ceiling. (Note that $x_j(\theta)$ is monotone increasing in $\theta$.)

**Lemma 4.8**

**Proof.** The result follows equation (4.19):

$$P_{\mu}(K) = \prod_{j=1}^{n} [\mu^{-1}_{i(j)}/\sum_{h=j}^{n} \mu^{-1}_{i(h)}] = \prod_{j=1}^{n} [c\mu^{-1}_{i(j)}/\sum_{h=j}^{n} c\mu^{-1}_{i(h)}] = P_{c\mu}(K),$$

which holds for all $c > 0$.

**Lemma 4.9**

**Proof.** By the closed-form solution of equation (4.19), the log-likelihood of a queue, $k$, is

$$\ln P(K = k) = \ln P(L_{i(1)} < \cdots < L_{i(n)}) = \sum_{j=1}^{n} \ln \mu^{-1}_{i(j)} - \sum_{j=1}^{n} \ln \left( \sum_{h=j}^{n} \mu^{-1}_{i(h)} \right).$$
Consider agent $i$ with position $p$ in the queue. Denote the sub-queue of all other agents by $K_{-i}$. Then fixing $p \in \{1, \ldots, n\}$ and a sub-queue realization $k_{-i}$, the log-likelihood of such a queue is

$$
\ln \mathbb{P}(K_i = p, k_{-i}) = \ln \mu_i^{-1} + \sum_{j \neq p}^{n} \ln \mu_{i(j)}^{-1} - \sum_{j=1}^{p} \ln \left( \mu_i^{-1} + \sum_{h \geq j, h \neq p} \mu_{i(h)}^{-1} \right)
$$

$$
- \sum_{j=p+1}^{n} \ln \left( \sum_{h=j}^{n} \mu_{i(h)}^{-1} \right).
$$

The marginal effect of an increase in $\mu_i^{-1}$ on the log-likelihood is

$$
\frac{\partial \ln \mathbb{P}(K_i = p, k_{-i})}{\partial (\mu_i^{-1})} = \mu_i - \sum_{j=1}^{p} \left[ \mu_i^{-1} + \sum_{h \geq j, h \neq p} \mu_{i(h)}^{-1} \right]^{-1}.
$$

In particular,

$$
\frac{\partial \ln \mathbb{P}(K_i = 1, k_{-i})}{\partial (\mu_i^{-1})} = \frac{\mu_i - 1}{\mu_i^{-1} + \sum_{j \neq i} \mu_j^{-1}} > 0;
$$

$$
\frac{\partial \ln \mathbb{P}(K_i = n, k_{-i})}{\partial (\mu_i^{-1})} = -\sum_{j=1}^{n-1} \left[ \mu_i^{-1} + \sum_{h=j}^{n-1} \mu_{i(h)}^{-1} \right]^{-1} < 0;
$$

and for $1 \leq p < n$, the difference

$$
\ln \mathbb{P}(K_i = p, k_{-i}) - \ln \mathbb{P}(K_i = p + 1, k_{-i}) = \left[ \sum_{h=p+1}^{n} \mu_{i(h)}^{-1} \right]^{-1} > 0.
$$

That is, the marginal effect is monotone decreasing, from positive to negative, along the queue position, $p$, of agent $i$. This monotonicity holds for any sub-queue realization $k_{-i}$. Therefore, for $c > 1$, there exists a $p^* \in \{1,\ldots, n-1\}$ such that $\mathbb{P}_\mu(K_i \leq k) = \sum_{p \leq k} \mathbb{P}_\mu(K_i = p) < \sum_{p \leq k} \mathbb{P}_{C,\mu}(K_i = p) = \mathbb{P}_{C,\mu}(K_i \leq k)$ if and only if $p \leq p^*$, i.e. $\mathbb{P}_{C,\mu}(K_i \leq k)$ is first-order stochastically dominated by $\mathbb{P}_\mu(K_i \leq k)$. The proof for the other half of the lemma is similar. \(\square\)

**Proposition 4.1**

**Proof.** Suppose the opposite, $\sum_i q_i < \bar{y}$, holds in equilibrium. Then by the quasi-concavity of $\pi(\cdot)$, $\hat{\pi}(y) > 0 = \hat{\pi}(\bar{y})$, $\forall 0 \leq y \leq \sum_i q_i$. Note that for any queue realization $k$, $Q_i^\leq + q_i \leq \sum_j q_j$ by construction (equation 4.1), and hence $\hat{\pi}(Q_i^\leq + q_i) > 0$. As the first-order condition (4.2) holds in equilibrium (lemma 4.1 guarantees the interior solution), this leads to a contradiction that the expectation of the product of strictly positive numbers equates
zero. Thus, the assumed inequality cannot hold in equilibrium. Instead, \( \sum_i q_i \geq \bar{y} \) and by quasi-concavity, \( \hat{\pi}(\sum_i q_i) \leq 0 = \hat{\pi}(\bar{y}) \).

Clearly, the “if and only if” statement holds true in the trivial case of \( n = 1 \). Consider \( n \geq 2 \) in what follows. First, the “only if” direction: Suppose there is an equilibrium with \( \sum_i q_i = \bar{y} \) (the equality holds) but there is no \( j \in \mathcal{N} \) such that \( \mathbb{P}(K_j = 1) = 1 \). Then for at least some \( j \in \mathcal{N} \), \( 0 < \mathbb{P}(K_j = 1) < 1 \) (at least someone might, but not almost surely, be the first in queue). Consider the first-order condition of such a market maker. A contradiction is derived from the first-order condition of such a market maker because the left-hand side is always strictly positive and it cannot equate zero: By the same argument as before, \( \hat{\pi}(y) > 0 \) for all \( 0 \leq y < \bar{y} \) and further, with probability \( \mathbb{P}(K_j = 1) \in (0, 1) \), \( \hat{\pi}(q_j) > 0 \). Second, the “if” direction: Suppose for some \( j \in \mathcal{N} \), \( \mathbb{P}(K_j = 1) = 1 \) but the equilibrium has the inequality strict, \( \sum_i q_i < \bar{y} \). This simply implies that the first-order condition for such an almost-surely first-in-queue market maker does not hold, as he can always improve his expected utility by increase his supply to \( \bar{y} \). Hence, it cannot be an equilibrium. 

**Proposition 4.2**

**Proof.** Differentiate the left-hand side of the first-order condition (4.2) with respect to \( q_j \) for some \( j \neq i \) to get \( \partial^2 \mathbb{E}\pi / (\partial q_i \partial q_j) = \mathbb{E}\left[ \mathbb{1}_{\{K_j < K_i\}} \cdot \hat{\pi}(Q_i^c + q_i) \right] \leq 0 \), by concavity of \( \pi(\cdot) \). This then establishes that \( q_i \) and \( q_j \) are strategic substitutes. Next, in equilibrium, the first-order condition (4.2) holds and it implies an implicit function of \( q_i = q_i^*(q_j) \). The substitution rate is the first order derivative of \( q_i \) with respect to \( q_j \), which by implicit function theorem is \( \partial q_i^*/\partial q_j = -\mathbb{E}[\mathbb{1}_{\{K_j < K_i\}} \hat{\pi}(Q_i^c + q_i)]/\mathbb{E}[\hat{\pi}(Q_i^c + q_i)] \). The partial derivative is negative because of the concavity, i.e. \( \hat{\pi}(\cdot) < 0 \). Note that there is at least one queue in which \( k_j > k_i \) and the indicator function becomes zero. Hence, the numerator is always strictly less negative than the denominator and the substitution rate is bounded by \((-1, 0)\). 

**Proposition 4.3**

**Proof.** The comparative static results can be shown through the implicit function of the zero-profit condition (4.6) where \( h = 0 \). Note that the zero-profit condition is essentially a first-order condition. Therefore, its partial derivative with respect to \( q(0) \) must be negative in equilibrium (as the second-order condition is necessary for optimality). It then remains to sign the partial derivatives with respect to \( \delta \) and to \( \eta \). The partial derivative with respect to \( \delta \) can be written as

\[
-\frac{\hat{\lambda}(\delta)}{\lambda(\delta)^2} \pi(\min\{\bar{y}, \delta q(0)/\eta\}) + \frac{1}{\delta \lambda(\delta) \eta} \delta q(0) \pi(\frac{\delta q(0)}{\eta}) \mathbb{1}_{\{\frac{\delta q(0)}{\eta} < \bar{y}\}}.
\]

If \( \delta q(0)/\eta \geq \bar{y} \), then the partial derivative is negative as \( \hat{\lambda}(\delta) = e^{\delta}/\tau > 0 \) by assumption 4.4 and hence proves \( \partial q(0)/\partial \delta < 0 \). If \( \delta q(0) < \bar{y} \), then note that \( \bar{y} \pi(y) < \pi(y) \) for
0 ≤ y ≤ \bar{y} because \( \pi(y) \) is concavely increasing by assumption 4.2. Then the above expression can be further evaluated as, with \( y := \delta q(0)/\eta \),

\[
- \frac{1}{\bar{\lambda}(\delta)^2} \frac{1}{\delta} \left[ \bar{\lambda}(\delta) \delta \pi(y) - \bar{\lambda}(\delta) y \pi(y) \right] < - \frac{1}{\bar{\lambda}(\delta)^2} \frac{1}{\delta} \pi(y) \left( \bar{\lambda}(\delta) \delta - \bar{\lambda}(\delta) \right) < 0,
\]

where the last inequality follows because \( \lambda(\delta) \) is convexly increasing for all \( \delta > 0 \).

Evaluate next the partial derivative with respect to \( \eta \) to get

\[
- \frac{\delta}{\eta^2} \frac{1}{\bar{\lambda}(\delta)} q(0) \pi \left( \frac{\delta}{\eta} q(0) \right) \mathbb{I}\{\frac{\delta}{\eta} q(0) < \bar{y}\},
\]

which is clearly non-positive as for all \( 0 ≤ y ≤ \bar{y} \), \( \pi(y) > 0 \). This completes the proof.

**Proposition 4.4**

**Proof.** Note that there are two components in the expression of \( \bar{\tau} \). Consider first the case where changes in the latencies do not affect \( \bar{h} \). Then

\[
\frac{\partial \bar{\tau}}{\partial \delta} = -\frac{1}{2} (\bar{h} + 1) \bar{h} + 1 - \eta \frac{\partial k_{h-1}^*}{\partial \delta} \approx -\frac{1}{2} (\bar{h} + 1) \bar{h} + 1 ≤ 0,
\]

where the approximation is valid for small \( \eta \). Similarly,

\[
\frac{\partial \bar{\tau}}{\partial \eta} = (\bar{h} + 1) - k_{h-1}^* - \eta \frac{\partial k_{h-1}^*}{\partial \eta} \approx (\bar{h} + 1) - k_{h-1}^* > 0,
\]

where the approximation is valid for small \( \eta \) and the inequality follows because \( k^* \) measures a fraction of market maker, whose maximal mass is 1. Therefore, if there is no jump of \( \bar{h} \), \( \bar{\tau} \) increases (decreases) in \( \eta \) (in \( \delta \), respectively).

Fixing \( \eta \), suppose \( \bar{h} \) jumps down from \( \bar{h} \) to \( \bar{h} - 1 \) as \( \delta \) increases to \( \delta + \epsilon \), where \( \epsilon \) is an infinitesimally small amount. Then by definition, the second component in \( \bar{\tau} \) is zero because \( k_{h-1}^* = \delta/\eta \) at the jump. The difference in \( \bar{\tau} \) is simply \( \bar{\tau}(\delta, \eta) - \bar{\tau}(\delta + \epsilon, \eta) = \eta \cdot (1 - \bar{h} \delta/\eta) \geq 0 \) because by construction \( 1 - h \delta/\eta \) is the mass of market makers who still have queuing uncertainty about their initial orders by round \( h \). This mass must be non-negative (otherwise the revision would have stopped already). Hence, as reaction latency \( \delta \) decreases, the order book stabilizes in a longer period of time.

Now fixing \( \delta \), suppose \( \bar{h} \) jumps up from \( \bar{h} \) to \( \bar{h} + 1 \) as \( \eta \) increases to \( \eta + \epsilon \), where \( \epsilon \) is an infinitesimally small amount. Then by definition, the second component in \( \bar{\tau} \) is zero because \( k_{h-1}^* = \delta/\eta \) at the jump. The difference in \( \bar{\tau} \) is simply \( \bar{\tau}(\delta, \eta) - \bar{\tau}(\delta, \eta + \epsilon) = -\eta \cdot (1 - (\bar{h} + 1) \delta/\eta) \leq 0 \). The inequality follows by the same argument as above. Hence, as transmission latency \( \eta \) decreases, the order book stabilizes in a shorter period of time. \( \square \)
Proposition 4.5

Proof. This proof shows the more general deep-in-book version of the proposition in section 4.6. From equation (4.16), the second-order derivative of \( \pi_j(y) \) can be computed as

\[
\pi_j(y) = -\rho(a_j - \nu([\theta_j] + \rho y))f([\theta_j] + \rho y).
\]

Suppose there exists some \( \bar{y} \) such that \( \pi_j(\bar{y}) = 0 \). Then

\[
0 = \pi_j(\bar{y}) = \int_{[\theta_j] + \rho \bar{y}}^{\infty} (a_j - \nu(\theta))f(\theta)d\theta < (a_j - \nu([\theta_j] + \rho \bar{y})) \cdot (1 - F([\theta_j] + \rho \bar{y})),
\]

or, \( a_j - \nu([\theta_j] + \rho y) = -\pi_j(\bar{y}) > 0 \). The inequality follows because \( \nu(\cdot) \) is assumed, as in Biais, Martimort, and Rochet (2000), to be increasing. By theorem M.C.4 of Mas-Colell, Whinston, and Green (1995), this implies the quasi-concavity of \( \pi_j(y) \). Hence, assumption 4.1 holds.

Similarly, suppose at some \( \hat{y} \) \( \hat{\pi}_j(\hat{y}) = 0 \), which implies \( a_j - \nu([\theta_j] + \rho \hat{y}) = 0 \) because \( \Theta \) has a continuous support and hence the density function is strictly positive on its support. Then it can be evaluated that the third-order derivative of \( \pi_j(y) \) at such a \( \hat{y} \) is \( \rho^2 \nu([\theta_j] + \rho y)f([\theta_j] + \rho y) > 0 \) (as \( \nu(y) > 0 \)). This implies that \( \hat{\pi}_j(y) \) is quasi-convex.

To prove assumption 4.2, suppose there exists some \( \bar{y}_j \) such that \( \pi_j(\bar{y}_j) = 0 \). By the quasi-concavity of \( \pi_j(\cdot), \pi_j(y) > 0 \) for all \( 0 \leq y < \bar{y}_j \). Then,

\[
0 \leq \hat{\pi}_j(y) = \int_{[\theta_j] + \rho y}^{\infty} (a_j - \nu(\theta))f(\theta)d\theta < (a_j - \nu([\theta_j] + \rho y))(1 - F([\theta_j] + \rho y))
\]

\[
= -\pi_j(y)\frac{f([\theta_j] + \rho y)}{(1 - F([\theta_j] + \rho y))},
\]

where the inequality “<” follows the monotonicity of \( \nu(\theta) \) and the last equality follows the expression of \( \hat{\pi}_j(\cdot) \). Rearrange the above inequality to get

\[
(4.20) \quad \hat{\pi}_j(y) < -\pi_j(y)\frac{f([\theta_j] + \rho y)}{1 - F([\theta_j] + \rho y)} < 0
\]

because the hazard rate is always positive. This shows that \( \pi_j(y) > 0 \) is sufficient for \( \pi_j(y) \) to be concave. ∎

Proposition 4.6

Proof. Immediately following lemma 4.6, the welfare maximizes at some \( y^* > \bar{y} \) because \( \hat{w}(\bar{y}) > 0 \). ∎
4.8. APPENDIX

Proposition 4.7

Proof. Define \( y := q_1 + q_2 \) as the aggregate book depth. In equilibrium, the first-order condition system (4.12) holds (the interior solution is guaranteed by lemma 4.1) and by chain-rule,

\[
\begin{align*}
\frac{\partial q_1}{\partial \alpha} &= \pi(y) \left[ \frac{1-\alpha}{\Delta} \pi(q_2) + 2\pi(y) \right] > 0 \\
\frac{\partial q_2}{\partial \alpha} &= -\pi(y) \left[ \frac{\alpha}{1-\alpha} \pi(q_1) + 2\pi(y) \right] < 0
\end{align*}
\]

where \( \Delta = \alpha \cdot (1-\alpha)\pi(q_1)\pi(q_2) + (1-\alpha)^2\pi(q_2)\pi(y) + \alpha^2\pi(q_1)\pi(y) > 0 \) is the determinant of the Jacobian matrix. To see why the partial derivatives (and \( \Delta \)) are so signed, note that by lemma 4.1, \( 0 \leq y = q_1 + q_2 \leq 2\bar{y} \) in equilibrium. Then by condition (4.13) \( \pi(y) < 0 \) because \( \pi(y) \) is quasi-convex and \( \pi(0) < 0 \) as implied by proposition 4.5. (Recall that \( \pi(y) < 0 \) in equilibrium because of liquidity overshoot.) The effect on aggregate book depth, \( y \), is

\[
\frac{\partial y}{\partial \alpha} = \frac{\partial q_1}{\partial \alpha} + \frac{\partial q_2}{\partial \alpha} = -\frac{\pi(y)}{\Delta} \left[ \frac{\alpha}{1-\alpha} \pi(q_1) - \frac{1-\alpha}{\alpha} \pi(q_2) \right].
\]

By symmetry, \( \partial y/\partial \alpha = 0 \) at \( \alpha = 1/2 \). It remains to show that \( \partial^2 y/\partial \alpha^2 < 0 \) at \( \alpha = 1/2 \), which is true as

\[
\frac{\partial^2 y}{\partial \alpha^2} \bigg|_{\alpha=1/2} = -4\pi(y)\pi(q_1)/\Delta < 0
\]

(note that \( q_1 = q_2 \) at \( \alpha = 1/2 \) by symmetry), making \( \alpha = 1/2 \) a local maximum. \( \square \)

Proposition 4.8

Proof. The proof goes in several steps. Step one shows how to derive the first-order condition (4.18). Note that for \( h \geq 1 \), \( \partial \pi_{j+h}(y)/\partial q_{i,j} = \frac{\partial}{\partial q_{i,j}} \left( \int_{\theta_j-h}^{\theta_j+h} \rho(y) (a_j + h - v(\theta)) f(\theta) d\theta \right) = \pi_{j+k}(y) - \pi_{j+k}(0) \). Hence,

\[
(4.21) \quad \frac{\partial}{\partial q_{i,j}} \left( \pi_{j+h}(y + q) - \pi_{j+h}(y) \right) = \pi_{j+h}(y + q) - \pi_{j+h}(y).
\]

The Bellman equation for price level \( j + 1 \) (see the \( j \)-th Bellman equation in equation 4.17) should hold in equilibrium: \( u_{i,j+1} = \mathbb{E} \left[ \pi_{j+1} \left( Q_{i,j+1} < q_i + q_{i,j+1} \right) - \pi_{j+1} \left( Q_{i,j+1} \leq q_i < q_{i,j+1} \right) \right] + u_{i,j+2} \). Optimality of the right-hand side requires the first order condition to hold:

\[
(4.22) \quad \mathbb{E} \pi_{j+1} (Q_{i,j+1} < q_i + q_{i,j+1}) + \frac{\partial u_{i,j+2}}{\partial q_{i,j+1}} = 0.
\]

Also, differentiate both sides of the Bellman equation with respect to \( q_{i,j} \) to get

\[
\frac{\partial u_{i,j+1}}{\partial q_{i,j}} = \frac{\partial}{\partial q_{i,j}} \mathbb{E} \left[ \pi_{j+1} \left( Q_{i,j+1} < q_i + q_{i,j+1} \right) - \pi_{j+1} \left( Q_{i,j+1} \leq q_i < q_{i,j+1} \right) \right] + \frac{\partial u_{i,j+2}}{\partial q_{i,j+1}} = -\mathbb{E} \pi_{j+1} (Q_{i,j+1} < q_i).
\]
where the second equality follows by substituting equation (4.21) with \( h = 1, y = Q^<_i, j+1 \), and \( q = q_{i,j+1} \), and equation (4.22). Finally, note that optimality of the Bellman equation (4.17) gives \( \mathbb{E}\hat{\pi}_j(Q^<_i + q_{i,j}) + \partial u_{i,j+1}/\partial q_{i,j} \). Substitute the above expression for \( \partial u_{i,j+1}/\partial q_{i,j} \) into the optimality, and the result is the first-order condition given in equation (4.18).

The second step derives an important properties about the expected profit function \( \pi_j(y) \). The quasi-concavity of \( \pi_j(y) \) and the quasi-convexity of \( \hat{\pi}_j(y) \) together imply the first property that if \( \hat{\pi}_j(y) \leq 0 \), then \( \hat{\pi}_j(y) \leq 0 \) for all \( y \geq 0 \). To see this, note that if the opposite is true, i.e. if there exists some \( \hat{y} \geq 0 \) such that \( \hat{\pi}_j(\hat{y}) > 0 \), then by continuity \( \hat{\pi}_j(y) \) must cross the horizontal axis at least once from below to above. The crossing point implies a local minimum of \( \pi_j(y) \), contradicting the quasi-concavity of \( \pi_j(y) \).

With the above results, the last step is to suppose \( \hat{\pi}_j(y_j) \geq 0 \), where \( y_j := \sum_i q_{i,j} \), and show this leads to a contradiction. Note that \( \{Q^<_i + q_{i,j}\} \leq y_j \) in equilibrium because there is non-zero probability for market maker \( i \)'s order not to queue in the last position. Given that \( \pi_j(y_j) \geq 0 \), therefore, \( \mathbb{E}\hat{\pi}_j(Q^<_i + q_{i,j}) = 0 \) by the inequality (4.20). Differentiate the first-order condition (4.18) on both sides with respect to \( q_{i,j} \): \( \mathbb{E}\hat{\pi}_j(Q^<_i + q_{i,j}) + \mathbb{E}\hat{\pi}_{j+1}(Q^<_i + q_{i,j}) = 0 \), or \( \mathbb{E}\hat{\pi}_{j+1}(Q^<_i + q_{i,j}) = -\mathbb{E}\hat{\pi}_j(Q^<_i + q_{i,j}) > 0 \), where the inequality follows the above argument. Then a contradiction follows:

\[
0 < \mathbb{E}\hat{\pi}_{j+1}(Q^<_i + q_{i,j}) < -\mathbb{E}\hat{\pi}_{j+1}(Q^<_i + q_{i,j}) \left[ \frac{f(|\theta_j| + Q^<_i q_{i,j})}{1 - F(|\theta_j| + Q^<_i q_{i,j})} \right] \\
< -\mathbb{E}\hat{\pi}_{j+1}(Q^<_i + q_{i,j}) \left[ \frac{f(|\theta_j|)q_{i,j}}{1 - F(|\theta_j|)} \right] < 0
\]

The second inequality follows inequality (4.20). The third inequality follows the increasing hazard rate of \( \Theta \). The last inequality follows the first-order condition (4.18) saying that \( \mathbb{E}\hat{\pi}_{j+1}(Q^<_i + q_{i,j}) = \mathbb{E}\hat{\pi}_j(Q^<_i + q_{i,j}) \), which is positive because \( \{Q^<_i + q_{i,j}\} \leq y_j \) and because of the quasi-convexity of \( \pi_j(\cdot) \). To conclude, this contradiction refutes the assumption that \( \hat{\pi}_j(y_{<j}) \geq 0 \) in equilibrium. Therefore, \( \hat{\pi}_j(y_{<j}) < 0 \), i.e. at each price level there is liquidity overshoot. Therefore, \( \hat{\pi}_j(y_j) < 0 \).

**Corollary 4.1**

**Proof.** Rewrite \( \partial y/\partial \alpha \) as

\[
\frac{\partial y}{\partial \alpha} = -\frac{\pi(y)}{\Delta} \pi(q_1) \alpha \left( \frac{\hat{\pi}(q_1)}{\pi(q_1)} - \frac{\hat{\pi}(q_2)}{\pi(q_2)} \right)
\]

by noting \( \pi(q_1)/\pi(q_2) = (1 - \alpha)^2/\alpha^2 \). Because \( \partial q_1/\partial \alpha > 0 \) and \( \partial q_2/\partial \alpha < 0 \) (see the proof of proposition 4.7), for \( \alpha \geq 1/2, q_1 \geq q_2 \) (that is, the faster market maker supplies
4.8. APPENDIX

more). Therefore, it suffices to sign the derivative of \( \hat{\pi}(x) / \pi(x) \) on \( x \in [0, \bar{y}] \): If \( \hat{\pi}(x) / \pi(x) \) is decreasing, then \( \partial y / \partial \alpha \leq 0 \) and \( y \) reaches its global maximum at \( \alpha = 1/2 \). Note that

\[
\frac{\partial}{\partial x} \left( \frac{\hat{\pi}(x)}{\pi(x)} \right) = \frac{1}{\pi(x)^2} \left( \hat{\pi}(x) \pi(x) - \pi(x)^2 \right)
\]

is negative if \( \hat{\pi}(x) \leq 0 \) (\( \pi(x) > 0 \) for \( x \in [0, \bar{y}] \) by quasi-concavity of \( \pi(\cdot) \)). That is, concavity of \( \hat{\pi}(x) \) on \( [0, \bar{y}] \) is a sufficient condition for the result.

Consider next the case where \( \pi(x) \) is not concave but only the first three order effects matter:

\[
\pi(x) = c_1 \cdot (x - \bar{y}) + c_2 \cdot (x - \bar{y})^2 + O((x - \bar{y})^3) \approx c_1 \cdot (x - \bar{y}) + c_2 \cdot (x - \bar{y})^2.
\]

Then it can be derived immediately that

\[
\frac{\partial}{\partial x} \left( \frac{\hat{\pi}(x)}{\pi(x)} \right) = \frac{1}{\pi(x)^2} \left( \hat{\pi}(x) \pi(x) - \pi(x)^2 \right) = -\frac{1}{\pi(x)^2} \left( c_2^2 \cdot (x - \bar{y})^2 + (c_1 + c_2 \cdot (x - \bar{y}))^2 \right) < 0.
\]

Hence, when the above approximation holds, \( \partial y / \partial \alpha \) is negative on \( \alpha \geq 1/2 \) and, by symmetry, positive on \( \alpha \leq 1/2 \). At \( \alpha = 1/2 \), \( y \) has the unique maximum.

\[\square\]

**Corollary 4.2**

**Proof.** The proof directly follows corollary 4.1, which implies the equilibrium book depth \( y \) is a one-to-one mapping from the effective queue distribution parameter, \( \alpha_R \), which in turn is a one-to-one mapping from \( \beta \), which governs the queue randomizer. Therefore, by choosing \( \beta \) (hence also \( R \)), the equilibrium depth can be adjusted to maximize welfare, which is a function in \( y \): \( w(y) = w(y(\alpha_R(\beta))) \). (Note that the optimal level \( \beta \) might be cornered.)

\[\square\]

**Corollary 4.3**

**Proof.** Suppose the equilibrium depth at some price level \( a_j \) is \( y_j > 0 \). By proposition 4.8, \( \hat{\pi}_j(y_j) = \int_{[\theta_j] + p y_j}^{\infty} (a_j - v(\theta)) f(\theta) d\theta < 0 \). Then the expected marginal profit of the first unit of limit orders at the next possible price, \( a_j + \rho \), is \( \int_{[\theta_j] + p y_j}^{\infty} (a_j + \rho - v(\theta)) f(\theta) d\theta \), which is negative for sufficiently small \( \rho \). Then by property 1) developed in the proof of proposition 4.8, the expected marginal profit at price \( a_j + \rho \) is always negative for all depth \( y \geq 0 \). Hence, no market maker is willing to post any order at this price level, leaving it a hole in the book.

\[\square\]

**Corollary 4.4**

**Proof.** In this proof, the superscript “\(*\)” indicates “with queuing uncertainty”, while the superscript “\(=\)” indicates “under the break-even condition of Sandås (2001)”. Consider the
expected marginal profit functions at price level $j$: $\dot{\pi}_j^*(y) = \int_{[\theta_j]^*}^{\theta_j^\infty} + \rho y (a_j^* - v(\theta)) f(\theta) d\theta$

and $\pi_j^\infty(y) = \int_{[\theta_j]^=}^{\theta_j^\infty} + \rho y (a_j^\infty - v(\theta)) f(\theta) d\theta$. In particular, $[\theta_j]^* = a_j^* - v_0 + \rho y_{<j}^*$ and $[\theta_j]^\infty = a_j^\infty - v_0 + \rho y_{<j}^\infty$. Decompose $\dot{\pi}_j^*(y)$ as

$$\dot{\pi}_j^*(y) = \int_{[\theta_j]^*}^{\theta_j^\infty} + \rho y (a_j^* - v(\theta)) f(\theta) d\theta + (a_j^* - a_j^\infty) (1 - F([\theta_j]^* + \rho y)).$$

When there is no holes in the order book, $a_j^* = a_j^\infty$, and the following inequality follows at equilibrium:

$$\pi_j^\infty(y_j^* + (a_j^* - a_j^\infty)/\rho + (y_{<j}^* - y_{<j}^\infty)) < \pi_j^*(y_j^*) < 0 = \pi_j^\infty(y_j^\infty).$$

By quasi-concavity of $\pi_j(\cdot)$, therefore, $y_j^* + (y_{<j}^* - y_{<j}^\infty) > y_j^\infty$, which simplifies to $y_{<j}^* > y_{<j}^\infty$. \qed


Ben-David, Itzhak, Francesco Franzoni, and Rabih Moussawi. 2012. “ETFs, Arbitrage, and
BIBLIOGRAPHY

Shock Propagation.” Working paper. 6


Race: Frequent Batch Auctions as a Market Design Response." Working paper. 80
CFTC and SEC. 2010a. “Findings regarding the market events of May 6, 2010.” Report, CFTC and SEC. 6, 7, 9, 11, 26, 28, 37, 38, 48, 49
———. 2010b. “Preliminary findings regarding the market events of May 6, 2010.” Report. 6


BIBLIOGRAPHY

The Journal of Finance 46 (2):733--746. 42


BIBLIOGRAPHY


Vives, Xavier. 2008. Information and Learning in Markets: The Impact of Market Mi-
crostructure. Princeton University Press. 51


NEDERLANDS SAMENVATTING

Summary in Dutch

This summary is based on the introduction in chapter 1. All credits shall go to Vincent van Kervel for this excellent translation.

Het doel van dit proefschrift is een bijdrage te leveren aan het begrip van financiële markten, en richt specifiek op de invloed van recente technologische ontwikkelingen op deze markten.

Technologische ontwikkelingen, tezamen met toezichthouders die concurrentie aansporen, hebben de organisatie van financiële markten drastisch verandert: De drempel om nieuwe beurzen op te zetten is sterk verlaagd over de jaren, en dit heeft de concurrentie tussen aandelenbeurzen aangezwengeld. De handel in aandelen is nu sterk verspreid over traditionele en nieuwe high-tech beurzen.

Daarnaast beïnvloeden technologische ontwikkelingen de handelstrategiën van individuele beleggers. De algoritmische handel maakt nieuwe beleggings-strategieën mogelijk, en dit trekt nieuwe partijen aan. Met extreme handelsnelheid domineren ze de markt, wat herkenbaar is aan opvallende data patronen. Andere beleggers passen zich hierop aan, waardoor een deel van de handel zich verplaatst naar nieuwe handelsplatformen zoals electronic communication networks, crossing networks en dark pools.

De drie hoofdstukken beschreven in dit proefschrift bestuderen de fricties veroorzaakt door deze technologische ontwikkelingen, en onderzoeken de gevolgen voor de liquiditeit en kwaliteit van de markt.

Hoofdstuk 2, gebaseerd op Menkveld and Yueshen (2014a), beschrijft een van de meest dramatische gebeurtenissen in de geschiedenis van de financiële markten: de 2010 Flash crash. Gedurende de Flash crash verdampte nine percent van de waarde van alle Amerikaanse beursgenoteerde bedrijven in minder dan twintig minuten. De studie gebruikt privacybeschermde data van de E-mini (S&P500 future) and SPY (S&P500 ETF), welke de individuele transacties bevat van een grote fundamental seller die naar verluid de Flash Crash veroorzaakte. Echter, het blijkt dat deze fundamental seller voorzichtig opereerde, en bijvoorbeeld weinig handelde gedurende de vrije val van de prijzen. Desondanks bleef ze verkopen op momenten dat de E-mini zeer laag stond, wat de prijs nog sterker deed dalen, en
maakte hierdoor excessieve verliezen. De grootte van de verliezen in deze twintig minuten komen overheen met ongeveer 25% van haar jaarlijkse bedrijfsopbrengst.

De empirische analyse toont aan dat electronische markten kwetsbaar kunnen zijn door ongelukkige interacties tussen beleggers. Dergelijke crashes verminderen het vertrouwen in financiële markten, en hebben zo grote gevolgen voor de gehele beleggingswereld.

Gebaseerd op Menkveld and Yueshen (2014b), Hoofdstuk 3 bestudeert de nieuwe middlemen, welke de vraag en aanbod in aandelen van de eind gebruikers intermediëren. Deze middlemen zijn doorgaans “high-frequency” market-makers, en zijn actief op vrijwel alle electronische beurzen. De contributie aan de literatuur is dat de interacties tussen middlemen wordt bestudeerd: Handelen ze ook met elkaar, en zo ja, hoe? Wat is de reden, en hoe beïnvloed dit andere beleggers?

Dit hoofdstuk beschrijft een model wat de invloed van middlemen op de welvaart van de gehele markt meet. Het bestudeert de positieve en negatieve effecten (aanbod van liquiditeit versus vermindering van de informatievoorziening). Het netto effect op welvaart hangt af van een parameter, welke aangeeft hoe lastig een middleman zijn huidige positie kan verkopen.

Ook wordt aangetoond dat een reguliere belegger “verward” raakt door marktactiviteit van de middlemen: tijdelijke prijsdruk tussen middlemen kan niet worden onderscheiden van fundamentele informatie over het aandeel. Bij tijdelijke prijsdruk is de reguliere belegger slechter af, aangezien deze door de prijsdruk slechtere prijzen krijgt. De voorspellingen van het model sluiten aan bij de gebeurtenissen van de Flash Crash (Hoofdstuk 2), en deze mechanismes lijken zeer relevant te zijn.

Hoofdstuk 4 is gebaseerd op Yueshen (2014). Het onderzoekt de optimale manier om limit orders naar de beurs te sturen, gegeven de frictie dat beleggers een verschillende reactie snelheid hebben. De reactie snelheid beslaat de tijd tussen de beslissing om te handelen en het moment dat een limit order op de beurs aankomt. Deze latencies kunnen random zijn, en hangen bijvoorbeeld af van de snelheid van de hardware en de internet verbinding. Latencies beïnvloeden de kans dat een belegger als snelste reageert op nieuws, en bepalen de volgorde in de wachtrij van limit orders in het order book. Onzekerheid in de volgorde in de wachtrij, zogeheten “queue-uncertainty”, heeft een sterke invloed op de winstgevendheid van een limit order.

Recente verbeteringen in hardware, aan de kant van zowel de beurzen als de beleggers, hebben het belang van queue-uncertainty alleen doen vergroten. Het model in Hoofdstuk 4 beschrijft deze frictie en analyseert de gevolgen ervan. Waar de huidige literatuur kijkt naar sequentiële strategiën, is dit model nieuw omdat alle beleggers simultaan handelen (wat leidt tot queue uncertainty). Deze aanpak genereert dynamische voorspellingen over hoe beleggers hun limit order strategie aanpassen naarmate ze hun locatie in de wachtrij observeren. Het model sluit aan op de literatuur van stabiele evenwichten in limit order markets, en kan een aantal empirische fenomenen voorspellen, zoals “schijn-liquiditeit”.
Ook geeft het nieuwe inzichten in hoe verschillende vormen van handel snelheid het evenwicht in een limit order book beïnvloed.
ACKNOWLEDGMENT

This thesis is tremendously indebted to the many who I shall bear, ever gratefully, in remembrance. It was my pleasure to have received help, guidance, encouragement, comments, and blessings during the years of my graduate study, beyond which I eventually realize there is also vivid life that I, too, have enjoyed, thanks to the many friends I had the honor to have been accompanied with.

Fortunate enough for my own sake, but not so much for the sake of completing this acknowledgment, the overwhelming gratefulness that I owe to these people is ever-growing, intimidating my ambition to “properly” write up this acknowledgment, arguably the most important (and the most read) chapter in a thesis. There are two main considerations. First, I fear that there would be people who should be acknowledged but were forgotten, however hard I try to ensure a complete coverage of my thankfulness. Second, it is an onerous task to sort out (i.e. to rank) who I should thank first and who in the second, so on and so forth. In view of these two challenges, I decide to compile an incomplete list of the people to whom my sincere thanks shall go in a random sequence.1

The only exception is dedicated to Albert J. Menkveld, my doctoral adviser, who trained, pushed, and supported me throughout the years with his passion for and insights in academic research. It is more than his training that I am acknowledging here. Above all, I feel obliged to his confidence (perhaps overconfidence) in my capability, his generosity in funding and in encouragement for my academic pursuit, and his cheering attitude towards scientific research. To quote, I thank Albert for having me on board the “research roller coaster”, experiencing the ups and downs together (and eventually up, up, and up).

For the truth unearthed and the falsifiable theories proposed in this thesis and for all other future academic achievements that I might fortuitously accomplish, I sincerely thank, in a randomized sequence with acknowledged incompleteness, Elvira Solji; Yvon Zhu for being “the lovely little red”; Dino Kong for yida-yida-yida; Yuyu Zeng; Andreas Park; Judith van Kronenburg for excellant secretary work and support at TI; Makoto Watanabe for helpful mock interview in preparation for my job market; Giovanni Cespa; Luyao He;

1 The declaration of the incompleteness is a preemptive effort to dissolve my future regrets, by which I am sure to be haunted, of failing to include some that should have been well acknowledged. The randomness of the list waives me from draining my brain power in figuring out the “rank”.

143
Thierry Foucault for the many useful comments on my research and for being on my thesis committee; Shmuel Baruch; Annika Tianle Jiang; Christina Carvalho for her excellent secretary support during my job market period; Roman Kozhan; Ka-fu Wong for introducing economics to me and for inspiring me to become a good teacher; Wing Wah Tham; Teodor Dyakov; Nan Yang; Bengt Holmström for hosting my MIT visit, for invaluable advices on research, and for writing a recommendation letter; Jean-Edouard Colliard; Jiangyu Ji for spilling over the unprecedented optimism and for redefining “deskmate”; Xiye Yang; Xiaoyu Shen; Lerby Ergun; “K.J.” Johannes Hagen for the days spent together in the “dungeon” of MIT Economics; Erkki Silde; Sarah Draus; Marius Zoican for helpful discussions on research and for peer pressure; Feng Dong for awing together the “unfathomably deep water”; István Barra for being cynic and for solving high-school (well, sometimes college) maths together; Rogier Potter van Loon; Ina Putter for excellent secretary work and support at VU; Jose L. Moraga for helping prepare my job market; Bernard Hosman for connecting my theoretical view of the financial market to the reality; Yu Gao and may my best wishes extend to Ning Liu and to their family; Bjoern Bruegmann for helpful mock interview session for my job market preparation; Tiantian Huang; Ye Mao; Te Bao; Kumar Venkataraman; Natalya Martynova; Haoxiang Zhu for very helpful discussions and comments on research and for a recommendation letter supporting me during the job market period; Michel van der Wel; Zhiling Wang for spilling over laughter, happiness, and optimism to me; Hedda Werkman for helpful discussions on research and for rediscovering my off-academic life; Zhaokun Zhang and may my best wishes extend to Yunqin Wang and to their family; Tong Wang; Chen Yao; Liting Zhou and may my best wishes extend to Xin Yang and to their family; Mark van Achter; Nico Dragt; Jérôme Dugast; Matthew Zaragoza-Watkins for almost nailing the “zombies”; Judy for, once upon a time, being the Judy; Wei Li; Xin Zhang; Yang Liu for, among the many other less worth mentioning, his excellence in dish cleaning; André Lucas for supporting me throughout my graduate study (from admitting me in TI to writing a recommendation letter during the job market period); Xiao Xiao; Peter Hoffmann; Lin Zhao for still be-friendning me the bastard, me the jerk, and me the asshole; Yun Dai; Yang Yang; Tiantian Huang; Elfie Bonke for excellent support, especially for housing, at VU; Simin He; Hao Zhang for being one to whom I can always talk; Javier H. Ospina for the years living on Prinsengracht and for our friendship; Johan Hombert; Marijke van der Horst; Evert Wipplinger; Konrad Raff for many helpful discussions on research, teaching, and preparation for the job market; Debby Tielman for excellent secretary work and support at VU; Ton Vorst; Ioanid Roșu for the coffee and the inspiring discussion in Cambridge, U.K.; my parents for letting their insubordinate only kid indulge in his own life; Guanyao Zhu and may my best wishes extend to Mia Fang and to their family; Xiaoming Cai; Ester van den Bragt for organizing SoFiE 2012 conference together; Milada Gadourek for
excellent secretary work and support at VU; Vincent van Kervel for spontaneous coffee breaks every now and then (and the marvelous discussions that always follow); Norman Seeger for, among many others, generously evaluating my teaching; Marcin Zamojski for the similar pride and snob that we share; Fannie Zhang for the first sight; Yiwei Zhang for the friendship that survives the time of a decade and the distance of a continent; Liping Lu; Janko Cizel; Frode Brevik; Sophie Moinas for her many thought-provoking, enlightening discussions on my research; Paul Muller; Arjen Siegmann for, among others, the you-know-what lunch topic that the “giant squid” always brings; Lu Liu; Boris van Leeuwen; Katya Malinova; Arianne de Jong for helping prepare the job market and for the excellent secretary service at TI; Zhenxing Huang; Xuedong Wang; Massimo Giuliodori for the helpful interview training for my job market; Patrick Verwijmeren; Dion Bongaerts; Kitty Jinglu Zhang; Elizaveta Mirgorodskaya; Jianwei Xin; Echo “Congee” Gu for helping with the apron; Peter Gautier for being on both my MPhil and PhD thesis committees; Keegan Zhaorong Li; and the many others that I—let me just repeat—shall bear ever gratefully in remembrance.
Corrigendum, if any, will be updated at http://goo.gl/CDI0Z1. Please direct suggestions, comments, or any spotted errors, typos, and mistakes to b@yueshen.me.
The Tinbergen Institute is the Institute for Economic Research, which was founded in 1987 by the Faculties of Economics and Econometrics of the Erasmus University Rotterdam, University of Amsterdam and VU University Amsterdam. The Institute is named after the late Professor Jan Tinbergen, Dutch Nobel Prize laureate in economics in 1969. The Tinbergen Institute is located in Amsterdam and Rotterdam. The following books recently appeared in the Tinbergen Institute Research Series:

538. T. BUSER. Essays in Behavioural Economics.


542. T. BAO. Experiments on Heterogeneous Expectations and Switching Behavior.


544. N. MOES. Cooperative Decision Making in River Water Allocation Problems.

545. P. STAKENAS. Fractional Integration and Cointegration in Financial Time Series.


547. J. ZENHORST. Macroeconomic Perspectives on the Equity Premium Puzzle.

548. B. PELLOUX. The Role of Emotions and Social Ties in Public On Good Games: Behavioral and Neuroeconomic Studies.


551. X. ZHANG. Modeling Time Variation in Systemic Risk.

552. H.R.A. KOSTER. The Internal Structure of Cities: the Economics of Agglomeration, Amenities and Accessibility.


554. J.L. MöHLMANN. Globalization and Productivity Micro-Evidence on Heterogeneous Firms, Workers and Products.

555. S.M. HOOGENDOORN. Diversity and Team Performance: A Series of Field Experiments.

556. C.L. BEHRENS. Product Differentiation in Aviation Passenger Markets: The Impact of Demand Heterogeneity on Competition.

557. G. SMRKOLJ. Dynamic Models of Research and Development.

558. S. PEER. The Economics of Trip Scheduling, Travel Time Variability and Traffic Information.
559. V. SPINU. Nonadditive Beliefs: From Measurement to Extensions.
561. M. VAN DUIJN. Location, Choice, Cultural Heritage and House Prices.
562. T. SALIMANS. Essays in Likelihood-Based Computational Econometrics.
563. P. SUN. Tail Risk of Equity Returns.
566. R.S. SCHOLTE. The Interplay Between Early-life Conditions, Major Events and Health Later in Life.
568. M. KILIÇ. Fundamental Insights in Power Futures Prices.
570. E.M.F. VAN DEN BROEK. Keeping up Appearances.
572. F.T. ZOUTMAN. A Symphony of Distributive Instruments.
573. M.J. GERRITSE. Policy Competition and the Spatial Economy.
574. A. OPSCHOOR. Understanding Financial Market Volatility.
579. E. RAVIV. Forecasting Financial and Macroeconomic Variables: Shrinkage, Dimension Reduction, and Aggregation.
580. J. TICHEM. Altruism, Conformism, and Incentives in the Workplace.
581. E.S. HENDRIKS. Essays in Law and Economics.
582. X. SHEN. Essays on Empirical Asset Pricing.
584. X. LI. Temporary Price Deviation, Limited Attention and Information Acquisition in the Stock Market.
585. Y. DAI. Efficiency in Corporate Takeovers.
587. A. SELIM. An Examination of Uncertainty from a Psychological and Economic Viewpoint.